A framework for identifying the most likely successful underprivileged tertiary bursary applicants

Renier Steynberg

Thesis presented in partial fulfilment of the requirements for the degree of
Master of (Industrial) Engineering
in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof JH van Vuuren
Co-supervisor: Mr DP Lötter

December 2016

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2016

Abstract

A number of *non-governmental organisations* (NGOs) are mandated to assist in the removal of financial barriers preventing underprivileged, prospective students from enrolling for tertiary studies, by managing the provision of bursaries to promising individuals. These NGOs are, however, often overwhelmed by the number of bursary applications they receive. In order to select the best applicants, very basic and sometimes unjustifiable methods involving weighted criteria are used in industry. A scientifically justifiable *decision support system* (DSS) framework is instead proposed in this thesis for aiding NGOs in this selection process.

This framework is capable of both predicting the tertiary study (success or failure) outcome and ranking of bursary applicants in terms of potential merit. The three main components of the framework are a *predictive component* (containing multiple statistical learning models in an ensemble manner which learn from past data and then make future outcome predictions in respect of new applicants), an *integration component* (which combines the predictions made by the aforementioned models into a single prediction for each applicant), and a *ranking component* (which produces a rank level for each applicant in addition to his or her combined prediction).

Examples of models that are included in the predictive component include *logistic regression*, *classification and regression trees*, *random forests*, the *C4.5 algorithm*, and *support vector machines*, while *majority voting* and *weighted majority voting* are examples of methodologies that may be included in the integration component. The working of the integration component is based on weighting the various model outputs according to their predictive accuracies in respect of a holdout set. Possible methodologies that may be included in the ranking component may be found within the realm of *multi-criteria decision analysis* techniques. Examples of these techniques are the *ELimination Et Choix Traduisant la REalite* III (ELECTRE III) and the *Preference Ranking Organisation METHod for Enrichment Evaluations* II (PROMETHEE II).

In order to demonstrate the practical use of the DSS framework, it is implemented in the context of sample data provided by two NGO industry partners. During an assessment of the performance of the DSS in this context, it is found that the accuracy of the combined success or failure predictions for applicants is superior to those of the individual models on a one-to-one comparison basis. It is also found that the average overall accuracy of the combined predictions surpasses that of the manual processes currently employed by the industry partners.

The sample data are further analysed for trends of interest and to identify those variables that seem to be best suited for predicting the tertiary success of prospective students. Surprising and perhaps counter-intuitive results are obtained, indicating that high school averages and subject marks are, in fact, negatively correlated to the eventual tertiary success of past students. This observation is likely due to better performing high school students gravitating to the more challenging, and potentially more prestigious, tertiary institutions, study fields, and qualification types.

Uittreksel

Verskeie *nie-regeringsorganisasies* (NROs) vervul die mandaat om finansiële struikelblokke uit die weg te ruim wat minder bevoorregte, voornemende studente daarvan weerhou om vir tersiêre studies in te skryf, deur die proses van beurstoekennings aan hierdie individue te bestuur. Hierdie NROs word egter dikwels oorval deur die ontvangs van menige beursaansoeke. In 'n poging om gepaste aansoekers vir finansiële steun te identifiseer, word baie eenvoudige en soms wetenskaplik onverantwoordbare metodes in die bedryf gebruik wat op die weging van kriteria berus. 'n Wetenskaplik verantwoordbare *besluitsteunstelsel* (BSS)-raamwerk word egter in hierdie tesis daargestel om NROs in hierdie moeilike seleksiebesluitnemingsproses by te staan.

Hierdie raamwerk is daartoe in staat om beide die tersiêre studie-uitkoms (sukses of mislukking) van beursaansoekers te voorspel en om hierdie aansoekers in volgorde van potensiële meriete te rangskik. Die drie hoofkomponente van hierdie raamwerk is 'n *voorspellingskomponent* (wat verskeie statistiese leermodelle op 'n ensemble-wyse inspan om uit historiese data te leer en dan voorspellings ten opsigte van nuwe beursaansoekers te maak), 'n *integrasiekomponent* (wat die voorspellings van die bogenoemde modelle tot 'n enkele voorspelling vir elke beursaansoeker kombineer), en 'n *rangorde-komponent* (wat buiten die voorspelling vir elke beursaansoeker ook 'n rangorde-vlak aan elke beursaansoeker toeken).

Voorbeelde van modelle wat by die voorspellingskomponent ingesluit word, sluit *logistiese regressie*, *klassifikasie- en regressiebome*, *lukrake woude*, die *C4.5 algoritme*, en *steunvektormasjiene* in, terwyl *meerderheidstemming* en *geweegde meerderheidstemming* voorbeelde is van metodologieë wat by die integrasiekomponent ingesluit kan word. Die werking van die integrasiekomponent berus op die weging van die onderskeie modelafvoere volgens die voorspellingsakkuraatheid van hierdie modelle in die konteks van 'n uithou-versameling. Moontlike metodologieë wat by die rangorde-komponent ingesluit kan word, spruit uit die studieveld van veelvuldige-kriteria besluitnemingsanalise en sluit *ELimination Et Choix Traduisant la REalite* III (ELECTRE III) en *Preference Ranking Organisation METHod for Enrichment Evaluations* II (PROMETHEE II) in.

Die praktiese toepasbaarheid van die BSS-raamwerk word gedemonstreer deur die stelsel op steekproefdata toe te pas wat deur twee NRO-nywerheidsvennote verskaf is. Gedurende 'n assessering van die BSS in hierdie konteks word daar bevind dat die akkuraatheid van die gekombineerde sukses- of mislukkingsvoorspellings vir beursaansoekers beter is as dié van die individuele modelle op 'n een-tot-een vergelykingsbasis. Daar word ook bevind dat die gemiddelde algehele akkuraatheid van die gekombineerde voorspellings dié van die huidige prosesse wat deur die nywerheidsvennote gebruik word, uitstof.

Die steekproefdata word verder analiseer om interessante neigings in die data sowel as veranderlikes te identifiseer wat goed gebruik kan word om die tersiêre sukses van voornemende studente te voorspel. Verbasende en moontlik teen-intuitiewe resultate word sodoende verkry wat trouens daarop dui dat gemiddeldes en vakpunte op hoërskool negatief met die uiteindelike tersiêre sukses van vorige studente korreleer. Hierdie waarneming kan moontlik daaraan toegeskryf word dat beter presterende hoërskoolleerders na meer uitdagende, en potensieel meer gesogte, tersiêre inrigtings, studievelde en kwalifikasietipes aangetrek word.

Acknowledgements

The author wishes to acknowledge the following people and institutions for their various contributions towards the completion of this work:

- J.P. and Hester Steynberg for their financial support and love over the past twenty-six years.
- Janneke Lötter (soon to be Steynberg) for her emotional support and understanding during the past two years.
- Anél Wesson for her immense effort of cleaning and providing sample data. Also, for her ever swift and friendly e-mail responses over the past two years.
- The two industry partners (who prefer to remain anonymous) which provided sample data for analysis in this thesis.
- Professor Jan van Vuuren and Danie Lötter for their monumental effort and exceptional guidance provided consistently over the past two years. Their impartation of knowledge include topics that reach far beyond the boundaries of these pages.
- The *Stellenbosch Unit for Operations Research in Engineering* (SUnORE) for the use of its excellent facilities within the Department of Industrial Engineering at Stellenbosch University. The academically stimulating environment fostered within the office significantly contributed to the productivity of the work of the author.
- The generous financial assistance provided by SUnORE in the form of a bursary.
- My fellow SUnORE members for the memories, constant encouragement, and other contributions in all forms that made the completion of this work possible.
- The One who provides imperishable riches, unceasing support, and continual guidance.

Table of Contents

Abstract iii

Uittreksel v

Acknowledgements vii

List of Reserved Symbols xiii

List of Acronyms xvii

List of Figures xix

List of Tables xxv

List of Algorithms xxix

1 Introduction 1

1.1 Background . . . 1

1.2 Thesis scope . . . 3

1.3 Thesis aim & methodology . . . 3

1.4 How this study relates to previous studies . . . 4

1.5 Thesis objectives . . . 6

1.6 Thesis organisation . . . 7

I Literature study 11

2 Information systems and data preparation 13

2.1 Raw data categories . . . 13

2.2 Data preparation . . . 15

2.3 Information systems . . . 18

2.4 System quality, testing, implementation and maintenance 27
2.5 Chapter summary 30

3 Statistical background, learning, regression, and classification 31
3.1 Statistical learning 32
3.2 Reasons for correlational relationships 46
3.3 Why is linear regression not appropriate for classification? 50
3.4 Logistic regression 53
3.5 Evaluation of a logistic regression model 56
3.6 Classification and regression trees 60
3.7 Random forests 67
3.8 The C4.5 algorithm 69
3.9 Support vector machines 72
3.10 Statistical and machine learning assumptions 78
3.11 Chapter summary 86

4 Ensemble learning and multi-criteria decision analysis 89
4.1 Ensemble learning 89
4.2 Integration strategies in ensemble learning 91
4.3 Multi-criteria decision making 93
4.4 Chapter summary 106

II Decision support system 107

5 Decision support system architecture 109
5.1 The proposed DSS framework 109
5.2 Data preparation 110
5.3 Model base 111
5.4 Chapter summary 113

6 Decision support system development 115
6.1 System development methodology and modelling tools 115
6.2 Study objectives and user requirements 116
6.3 Analysis of the system needs 117
6.4 Design of the system 120
6.5 Building of the system 120
6.6 Quality assurance, verification and validation 137

6.7 Chapter summary 142

III Case studies 143

7 Case study data presentation and preparation 145

7.1 Target population 146

7.2 Discussion of the main industry partner sample data 146

7.3 Case Study A aim and variable selection 152

7.4 Case Study B aim and variable selection 154

7.5 Discussion of the secondary NGO sample data 158

7.6 Case Study C aim and variable selection 160

7.7 Case Study D aim 162

7.8 Chapter summary 162

8 Case Study analyses 163

8.1 Case Study A 163

8.2 Roadmap for assessing variable importance 171

8.3 Case Study B 174

8.4 Case Study C 219

8.5 Case Study D 238

8.6 Chapter summary 242

IV Conclusion 243

9 Thesis summary & appraisal 245

9.1 Thesis summary 245

9.2 Appraisal of thesis contributions 247

10 Further work 251

10.1 The variable importance analyses and trend investigations 251

10.2 The predictive component of the proposed DSS 252

10.3 The integration strategy component of the proposed DSS 252

10.4 The ranking component of the proposed DSS 253

10.5 Performance assessments of proposed the DSS 253

10.6 The graphical user interface 253

10.7 The DSS framework 254

References **255**

V Appendices **269**

A Communication registry **271**

B R libraries employed **275**

C Additional graphs and tables of Case Study B **277**

List of Reserved Symbols

Symbols in this thesis conform to the following font conventions		
$\boldsymbol{a}$, $\boldsymbol{A}$	Symbol denoting a **vector or matrix**	(Boldface symbols in mathematics font)
$\mathcal{A}$	Symbol denoting a **set**	(Calligraphic capitals)

Symbol	Meaning
General	
$\mathcal{N}$	Set of observations in sample data
n	Number of observations or alternatives in sample data
m	Number of independent variables
K	Number of output classes
Linear and logistic regression	
X_i	Input value of observation i
$X_{j,i}$	The j-th input variable value of observation i
Y_i	Output value of observation i
R_i	Residual for the pair (X_i, Y_i)
β_0	Intercept coefficient
β_i	Slope coefficient of observation i
ϵ_i	True error of observation i
r	Pearson's r coefficient
H_0	Null hypothesis
H_a	Alternative hypothesis
α	Significance level
$P(\cdot)$	Probability of an event $(\cdot)$ occurring
G	Goodness of fit index
W	Wald statistic
c	Pre-determined critical value associated with Wald statistic
SE_i	Standard error estimate of observation i
Classification and regression trees and Random forests	
p_{tk}	Proportion of observations of class k in CART node t
T_{root}	Root node of a tree built from sample data
$T_{\max}$	Maximum tree built from sample data
I	Total number of parent nodes in a tree
t_p	General parent node
t_ℓ	General left child node
t_r	General right child node

t_p^i	Parent node i
t_ℓ^i	Left child node of parent node i
t_r^i	Right child node of parent node i
x'_j	Optimal splitting value of variable x_j
$i(t_p)$	Impurity function of t_p
$i(t_c)$	Impurity of t_ℓ and t_r combined
$\Delta i(t)$	Change in impurity of node t
P_ℓ	Probability of left node classification
P_r	Probability of right node classification
T	Classification and regression tree
$R(T)$	Misclassification rate or re-substitution error of T
$\lvert\tilde{T}\rvert$	Number of terminal nodes in T
λ	Complexity parameter
$\lambda\lvert\tilde{T}\rvert$	Measure of the complexity of T
$C_\lambda(T)$	Cost complexity of T
$T(\lambda)$	The smallest subtree pruned from T_{max} for a fixed value of λ
T_t	Branch of T rooted at node t
$C_\lambda(t)$	Cost complexity of node t
$C_\lambda(T_t)$	Cost complexity of branch of T rooted at node t
λ_{cut}	The λ-value at which $C_\lambda(t)$ and $C_\lambda(T_t)$ are equal
T_k	Smallest minimising subtree of T_{max}, starting at $k = 1$ and $\lambda = 0$
T_{t_k}	Branch pruned from T_k at node t
$g(t_k)$	λ_{cut}-value for each node t_k of the tree T_k
$g(\bar{t}_k)$	The weakest link node ($\min g(t_k)$)
ntree	Number of random tree built for random forests
mtry	Number independent variables considered as possibly splitting variables
$\mathbf{prox}(\boldsymbol{i},\boldsymbol{j})$	An $n \times n$ proximity matrix

The C4.5 algorithm	
t_s	Child node s split from t_p
c	Number of child nodes split from t_p
h	Number of categories present in the best qualitative independent variable
$\mathrm{freq}(C_j, S)$	Number of observations of a subset S of the sample data that belong to class j
J	Total number of classes considered
S	Particular subset S of the sample data

Support vector machines	
H	Hyperplane
B_1, B_2	Boundaries lines
M	Largest margin of separation between B_1 and B_2
$\boldsymbol{w}$	Vector perpendicular to H
b	Perpendicular distance from the origin to H
y_i	Outcome class of observation i
α_i	Lagrange multiplier of observation i
ξ_i	Slack variable of observation i
C	Upper bound on $\alpha_1, \alpha_2, \ldots, \alpha_n$
$k(\boldsymbol{x}_i, \boldsymbol{x}_j)$	Kernel function value of the inner product $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$
σ	Parameter of the Gaussian radial basic function kernel

Multi-criteria decision analysis (MCDA)	
$\mathcal{A}$	Set of alternatives in an MCDA model
$\mathcal{G}$	Set of criteria in an MCDA model
u	Number of criteria in an MCDA model
$\boldsymbol{E}$	An $n \times u$ performance matrix
w_j	Weight of criterion j
$(a\,\boldsymbol{P}\,b)$	Preference situation: alternative a is strongly preferred to alternative b
$(a\,\boldsymbol{Q}\,b)$	Weak preference situation: alternative a is weakly preferred to alternative b
$(a\,\boldsymbol{I}\,b)$	Indifference situation: indifferent between alternative a and alternative b
$(a\,\boldsymbol{R}\,b)$	Incomparability situation: alternative a is incomparable with alternative b
$(a\,\boldsymbol{S}\,b)$	Outranking relation: alternative a outranks alternative b
q	Indifference threshold
p	Preference threshold
v	Veto threshold
$S(a, b)$	Credibility index of the allegation $(a\,\boldsymbol{S}\,b)$
$c_j(a, b)$	Concordance index value of alternative a *versus* alternative b for criterion j
$d_j(a, b)$	Discordance index value of alternative a *versus* alternative b for criterion j
$g_j(i)$	Evaluation of alternative i with respect to criterion j
$Q(i)$	Qualification of alternative i
Z_1	Descending distillation levels of alternatives
Z_2	Ascending distillation levels of alternatives

Other	
$^\circ C$	Degrees Celsius
K	Kelvin

List of Acronyms

AUC	Area Under the Curve
CART	Classification and Regression Trees
CASE	Computer Assisted Software Engineering
CI	Confidence Interval
CPUT	Cape Peninsula University of Technology
CSV	Comma Separated Values
CUT	Central University of Technology
CV	Cross-validation
CVM	Cross-validation Majority
DUT	Durban University of Technology
DW	Durbin-Watson coefficient
EC	Eastern Cape
ELECTRE	*ELimination Et Choix Traduisant la REalite*
ERD	Entity-Relationship Diagrams
FS	Free State
GP	Gauteng Province
GUI	Graphical User Interface
HCI	Human-Computer Interaction
IDE	Integrated Development Environment
IS	Information System
KZN	KwaZulu-Natal
LP	Limpopo Province
MADM	Multi-attribute Decision Making
MCDA	Multi-criteria Decision Analysis

MCDM	Multi-criteria Decision Making
MODM	Multi-objective Decision Making
MS	MicroSoft
NDA	Non-disclosure Agreement
NGO	Non-governmental Organisation
NMMU	Nelson Mandela Metropolitan University
NW	North West
NWU	North West University
OOB	Out-of-Bag
OOM	Object-orientated Methodology
OR	Odds Ratio
ROC	Receiver Operating Curve
SDLC	Systems Development Life Cycle
SDM	Systems Development Methodology
SMHSU	Sefako Makgatho Health Sciences University
SQL	Structured Query Language
SU	Stellenbosch University
SVM	Support Vector Machine
TUT	Tshwane University of Technology
UCT	University of Cape Town
UFS	University of the Free State
UI	User Interface
UJ	University of Johannesburg
UKZN	University of KwaZulu-Natal
UML	Unified Modelling Language
UP	University of Pretoria
UWC	University of the Western Cape
VIF	Variance Inflation Factor
VIM	Variable Importance Measure
WC	Western Cape
XML	Extensible Markup Language

List of Figures

1.1 2014 Gini Index world map 2

1.2 Unequal living conditions in two suburbs in South Africa 2

2.1 Data variable types 14

2.2 The four-level pyramid model 19

2.3 An iterative UI design and evaluation process 21

2.4 Framework for decision making using multiple techniques or models 22

2.5 Seven phases of the systems development life cycle 24

2.6 The traditional waterfall design methodology 25

2.7 Activity diagram elements used in the OOM 26

2.8 Four steps of system testing 28

3.1 The quantities involved in simple linear regression 34

3.2 Simple linear regression — residuals *versus* true errors 35

3.3 Perfect positive correlation illustrative example 36

3.4 High positive correlation illustrative example 36

3.5 Low positive correlation illustrative example 36

3.6 No correlation illustrative example 36

3.7 Perfect negative correlation illustrative example 36

3.8 High negative correlation illustrative example 36

3.9 Low negative correlation illustrative example 36

3.10 Perfect negative correlation illustrative example 36

3.11 Three estimates of income as a function of years of education and seniority . . . 42

3.12 Partition of learning and validation subsets 45

3.13 The notion of K-fold cross-validation 45

3.14 The notion of bootstrapping 46

3.15 Scatter plot of individuals who did and did not default on credit card payments . 51

3.16 Probabilities of individuals defaulting, by applying linear regression 53

3.17 Probabilities of individuals defaulting, by applying logistic regression 54
3.18 Parent and child nodes in a CART 61
3.19 Misclassification rate *versus* tree size 64
3.20 Cost complexity of a single node t and its branch T_t as λ increases 65
3.21 SVM fundamentals 73
3.22 SVM margin of separation 73
3.23 SVM support vectors 73
3.24 Maximising the margin of separation in an SVM 74
3.25 Linearly non-separable SVMs 76
3.26 Examples of homoscedasticity and heteroscedasticity 80
3.27 Examples of linear and curvilinear relationships 81

4.1 An ensemble method which combines output from u different models 90
4.2 The first stage of Mota's typological tree 104
4.3 The second stage of Mota's typological tree 104
4.4 The third stage of Mota's typological tree 105
4.5 The fourth stage of Mota's typological tree 105

5.1 Overall working of the proposed DSS 110
5.2 Working of the data preparation component 111
5.3 Working of the assessing data quality and performing data cleaning component . 111
5.4 Working of the model base component 112

6.1 The seven phases of the SDLC 116
6.2 Use case model for the main industry partner 118
6.3 Activity diagram of the process followed by the main industry partner 119
6.4 The four stages of Mota's typology tree applied to the current study 123
6.5 Software considerations in the development of the DSS concept demonstrator . . 125
6.6 Files of input data are selected for importation into the system 126
6.7 Selection of a unique identifier, input variables, and an output variable 127
6.8 Number of data entries to be deleted due to missing values 128
6.9 Initial selection of base models available to the user 129
6.10 Identification of outliers and influential observations 129
6.11 Statistical learning model assumptions overview 130
6.12 Independence of residuals 130
6.13 Multicollinearity 131
6.14 Outliers and influential values 131

6.15 Minimum sample size 131
6.16 Linearity of the logit 132
6.17 Selecting the final base models 132
6.18 Prediction and ranking of test set data (top) 133
6.19 Prediction and ranking of test set data (middle) 133
6.20 Prediction and ranking of test set data (bottom) 133
6.21 Prediction and ranking of new alternatives (top) 134
6.22 Prediction and ranking of new alternatives (middle) 134
6.23 Prediction and ranking of new alternatives (bottom) 134
6.24 Discrepancy between the weighted prediction and ranking of new alternatives . . 134
6.25 Example of an error message presented to the user 135
6.26 Three examples of confirmation and error messages displayed to the user 136
6.27 An example of the error message shown to the user if singularity is found 136
6.28 An example of the progress bar 136
6.29 First GUI flow from the perspective of the user 140
6.30 Second GUI flow from the perspective of the user 141

7.1 Target population applicable to this study 146
7.2 Outlier detection performed on the sample data of Case Study A 154

8.1 Random forest variable importance for the model of Case Study A 166
8.2 Second random forest variable importance scores for the model of Case Study A 166
8.3 Interpretation key for the CART plot nodes 167
8.4 CART plot for the model of Case Study A 167
8.5 Graduation rate *versus* qualification type in Case Study A 168
8.6 Graduation rate *versus* Grade 11 November averages in Case Study A 168
8.7 Random forest variable importance for Combination 1 of Case Study B 193
8.8 Random forest variable importance for Combination 4 of Case Study B 193
8.9 Random forest variable importance for Combination 5 of Case Study B 193
8.10 Random forest variable importance for Combination 6 of Case Study B 194
8.11 Random forest variable importance for Combination 7 of Case Study B 194
8.12 Second random forest variable importance scores for Combinations 27 and 28 . . 195
8.13 Random forest variable importance for Combinations 27 and 28 of Case Study B 195
8.14 Second random forest variable importance for Combinations 29 and 30 196
8.15 Random forest variable importance for Combinations 29 and 30 of Case Study B 197
8.16 Random forest variable importance for Combination 32 of Case Study B 197

8.17 Random forest variable importance for Combinations 33 and 34 of Case Study B 198
8.18 Second random forest variable importance for Combinations 33 and 34 198
8.19 CART plot for Combination 3 of Case Study B 199
8.20 CART plot for Combination 4 of Case Study B 200
8.21 CART plot for Combination 5 of Case Study B 201
8.22 CART plot for Combination 6 of Case Study B 201
8.23 CART plot for Combination 7 of Case Study B 202
8.24 CART plot for Combination 8 of Case Study B 202
8.25 CART plot for Combination 9 of Case Study B 203
8.26 CART plot for Combination 10 of Case Study B 203
8.27 CART plot for Combination 11 of Case Study B 204
8.28 CART plot for Combination 17 of Case Study B 204
8.29 CART plot for Combination 23 of Case Study B 205
8.30 CART plot for Combination 25 of Case Study B 205
8.31 CART plot for Combinations 27 and 28 together of Case Study B 206
8.32 CART plot for Combinations 29 and 30 together of Case Study B 207
8.33 CART plot for Combination 31 of Case Study B 207
8.34 CART plot for Combination 32 of Case Study B 208
8.35 CART plot for Combinations 33 and 34 combined of Case Study B 208
8.36 Graduation rate as a function of race in Case Study B 209
8.37 Graduation rate as a function of gender in Case Study B 209
8.38 Graduation rate as a function of study region in Case Study B 210
8.39 Graduation rate as a function of tertiary institution in Case Study B 210
8.40 Graduation rate as a function of study field in Case Study B 210
8.41 Graduation rate as a function of qualification type in Case Study B 211
8.42 Graduation rate *versus* Grade 11 November average in Case Study B 212
8.43 Graduation rate *versus* Grade 12 June average in Case Study A 212
8.44 Graduation rate *versus* Grade 12 November average in Case Study A 212
8.45 Graduation rate *versus* Grade 12 Accounting marks in Case Study B 212
8.46 Graduation rate *versus* Grade 12 English (2nd) marks in Case Study B 213
8.47 Graduation rate *versus* Grade 12 Mathematics marks in Case Study B 213
8.48 Graduation rate *versus* Grade 12 Physical science marks in Case Study B 213
8.49 Random forest variable importance for Combination 1 of Case Study C 226
8.50 Random forest variable importance for Combination 2 of Case Study C 227
8.51 Random forest variable importance for Combination 3 of Case Study C 227

8.52 Random forest variable importance for Combination 4 of Case Study C 227
8.53 Random forest variable importance for Combination 8 of Case Study C 227
8.54 Random forest variable importance for Combination 9 of Case Study C 228
8.55 Second random forest variable importance for Combination 9 of Case Study C . 228
8.56 Random forest variable importance for Combination 10 of Case Study C 229
8.57 Second random forest variable importance for Combination 10 of Case Study C . 229
8.58 CART plot for Combination 2 of Case Study C 230
8.59 CART plot for Combination 3 of Case Study C 230
8.60 CART plot for Combination 4 of Case Study C 231
8.61 CART plot for Combination 5 of Case Study C 231
8.62 CART plot for Combination 6 of Case Study C 232
8.63 CART plot for Combination 7 of Case Study C 232
8.64 CART plot for Combination 8 of Case Study C 233
8.65 CART plot for Combination 9 of Case Study C 233
8.66 CART plot for Combination 10 of Case Study C 234
8.67 Graduation rate as a function of tertiary institution in Case Study C 234
8.68 Graduation rate as a function of gender in Case Study C 235
8.69 Graduation rate as a function of study field in Case Study C 235
8.70 Graduation rate as a function of race in Case Study C 235
8.71 Graduation rate *versus* Grade 12 Mathematics mark in Case Study C 236
8.72 Graduation rate *versus* Grade 12 Physical science mark in Case Study C 236
8.73 Graduation rate *versus* Grade 12 English mark in Case Study C 236

C.1 Graduation rate *versus* Grade 12 Afrikaans (1st) marks in Case Study B 277
C.2 Graduation rate *versus* Grade 12 Afrikaans (2nd) marks in Case Study B 278
C.3 Graduation rate *versus* Grade 12 Agricultural science marks in Case Study B . . 278
C.4 Graduation rate *versus* Grade 12 Business economics marks in Case Study B . . 278
C.5 Graduation rate *versus* Grade 12 Economics marks in Case Study B 278
C.6 Graduation rate *versus* Grade 12 Geography marks in Case Study B 279
C.7 Graduation rate *versus* Grade 12 History marks in Case Study B 279
C.8 Graduation rate *versus* Grade 12 isiXhosa marks in Case Study B 279
C.9 Graduation rate *versus* Grade 12 isiZulu marks in Case Study B 279
C.10 Graduation rate *versus* Grade 12 Life science marks in Case Study B 280
C.11 Graduation rate *versus* Grade 12 Mathematics literacy marks in Case Study B . 280
C.12 Graduation rate *versus* Grade 12 Setswana marks in Case Study B 280

List of Tables

2.1 Aid for selecting between the waterfall, agile, and OOM methodologies 27

3.1 Possible outcomes of hypothesis tests 37

3.2 Example of OR and 95% CI boundaries 59

3.3 Example OR results for (3.10) 59

3.4 Classification table for comparing predictive outcomes *versus* actual outcomes . . 60

3.5 Inner product kernel function examples 78

3.6 Assumptions associated with various methods of statistical analysis 87

4.1 Example of static weighting 92

4.2 Example of dynamic weighting 92

4.3 Metacriteria for distinguishing between MODM and MADM methods 93

4.4 Two mixed MCDA methods 95

4.5 Various elementary MCDA methods 96

4.6 Various single synthesising criterion MCDA methods 96

4.7 Various outranking MCDA methods 97

4.8 The MCDA methods of Tables 4.4–4.7 *versus* MCDA selection guidelines 103

6.1 MCDA methods suitable for use in the current study 124

6.2 Possible progress bar messages displayed to the user in tab seven 136

7.1 Fields in the sample data used in Case Studies A and B 147

7.2 Gender options applicable to Case Studies A and B 147

7.3 Race options applicable to Case Studies A and B 147

7.4 South African provinces applicable to Case Studies A and B 148

7.5 Tertiary institutions supported by the main industry partner 148

7.6 Academic statuses applicable to Case Studies A and B 149

7.7 Study fields applicable to Case Studies A and B 150

7.8 Qualification types applicable to Case Studies A and B 150

7.9 High school subjects applicable to Case Studies A and B 151
7.10 Recommendation score criteria applicable to Case Study A 152
7.11 Erroneous data and corrective strategies of sample data of Case Study A 153
7.12 Independent variable corrected due to erroneous values in Case Study A 154
7.13 Additional tertiary institutions in the sample data of Case Studies A and B . . . 155
7.14 Students deleted due to incorrect dependent variable labels in Case Study B . . . 157
7.15 Error values in and corrective strategies in sample data for Case Study B 157
7.16 Numbers of independent variables corrected in the sample data of Case Study B 158
7.17 Missing entries for independent variables in the sample data of Case Study B . . 159
7.18 Fields in the sample data used for Case Study C 160
7.19 Gender options applicable to Case Study C 160
7.20 Race options applicable to Case Study C 160
7.21 Tertiary institutions supported by the secondary industry partner 160
7.22 Academic statuses applicable to Case Study C 160
7.23 Error values in and corrective strategies of sample data for Case Study C 161
7.24 Independent variables corrected in the sample data of Case Study C 162
7.25 Missing entries for independent variables of the sample data for Case Study C . . 162

8.1 The p-values, $\hat{\beta}$-values, ORs, and their CIs for the model of Case Study A 165
8.2 Independent variables and reference classes for the sample data of Case Study B 174
8.3 Available and required observations for combinations of Case Study B 176
8.4 Assumptions results of the sample data of Case Study B 178
8.5 Log likelihood p-values for each combination of Case Study B 179
8.6 Available and required observations for additional combinations of Case Study B 179
8.7 Assumption results for additional combinations of Case Study B 179
8.8 Log likelihood p-values for each additional combination of Case Study B 180
8.9 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 1 of Case Study B . 180
8.10 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 3 of Case Study B . 180
8.11 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 4 of Case Study B . 181
8.12 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 5 of Case Study B . 181
8.13 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 6 of Case Study B . 182
8.14 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 7 of Case Study B . 183
8.15 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 8 of Case Study B . 183
8.16 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 9 of Case Study B . 183
8.17 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 10 of Case Study B . 184
8.18 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 11 of Case Study B . 184

8.19 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 17 of Case Study B . 184

8.20 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 23 of Case Study B . 184

8.21 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 25 of Case Study B . 185

8.22 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 27 of Case Study B . 185

8.23 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 28 of Case Study B . 186

8.24 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 29 of Case Study B . 187

8.25 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 30 of Case Study B . 188

8.26 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 31 of Case Study B . 190

8.27 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 32 of Case Study B . 190

8.28 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 33 of Case Study B . 190

8.29 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 34 of Case Study B . 191

8.30 Students' Grade 12 November marks *versus* specific tertiary institutions 215

8.31 Students' Grade 12 November marks *versus* specific field of study 215

8.32 Students' Grade 12 November marks *versus* specific qualification type 216

8.33 Students' Grade 12 November marks *versus* specific study region 216

8.34 Average weighted prediction accuracy of combinations of Case Study B 217

8.35 Independent variables and reference classes for sample data of Case Study C . . 220

8.36 Number of available and required observations for combinations of Case Study C 220

8.37 Outliers, influential observations, and assumptions results of Case Study C . . . 221

8.38 Log likelihood p-values for each combination of Case Study C 221

8.39 Available and required observations for additional combinations of Case Study C 221

8.40 Assumption results for additional combinations of Case Study C 222

8.41 Log likelihood p-values for each additional combination of Case Study C 222

8.42 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 1 of Case Study C . 223

8.43 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 2 of Case Study C . 223

8.44 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 3 of Case Study C . 223

8.45 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 4 of Case Study C . 224

8.46 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 5 of Case Study C . 224

8.47 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 6 of Case Study C . 224

8.48 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 7 of Case Study C . 224

8.49 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 8 of Case Study C . 224

8.50 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 9 of Case Study C . 225

8.51 The p-values, $\hat{\beta}$-values, ORs, and their CIs for Combination 10 of Case Study C . 226

8.52 Average weighted prediction accuracy of combinations of Case Study C 237

8.53 Weighted prediction accuracy for ten combinations of Case Study B 239

8.54 Weighted prediction accuracy for two combinations of Case Study C 239

A.1 Registry of communication with main industry partners 271

B.1 Main R libraries and functions used during creation of the DSS 276

C.1 Students' Grade 11 November marks *versus* specific tertiary institutions 281
C.2 Students' Grade 11 November marks *versus* specific study fields 281
C.3 Students' Grade 11 November marks *versus* specific qualification types 282
C.4 Students' Grade 11 November marks *versus* specific study regions 282
C.5 Students' Grade 12 Mathematics marks *versus* specific tertiary institutions . . . 282
C.6 Students' Grade 12 Mathematics marks *versus* specific study fields 283
C.7 Students' Grade 12 Mathematics marks *versus* specific qualification types 283
C.8 Students' Grade 12 Mathematics marks *versus* specific study regions 283
C.9 Students' Grade 12 Physical science marks *versus* specific tertiary institutions . . 284
C.10 Students' Grade 12 Physical science marks *versus* specific study fields 284
C.11 Students' Grade 12 Physical science marks *versus* specific qualification types . . 285
C.12 Students' Grade 12 Physical science marks *versus* specific study regions 285

List of Algorithms

3.1 Determining the tree T_1 65

3.2 Computing the nested sequence of trees $T_{k+1} \supset T_{k+2} \supset \cdots \supset T_{\mathrm{root}}$ 66

3.3 C4.5 tree construction algorithm 70

CHAPTER 1

Introduction

Contents

1.1 Background 1
1.2 Thesis scope 3
1.3 Thesis aim & methodology 3
1.4 How this study relates to previous studies 4
1.5 Thesis objectives 6
1.6 Thesis organisation 7

1.1 Background

In 1912, the Italian statistician Corrado Gini developed a measure of statistical dispersion known as the *Gini Index* [39, p. 421]. Today this index is widely used as the international standard for measuring the income and wealth distribution of a country's residents. Figure 1.1 contains a map of the Gini Index of income inequality distribution per country according to data obtained from the World Bank, as published in July 2014 [224]. From this map it is clear that South Africa has one of the greatest divides between rich and poor individuals in the world. This divide is also clearly visible in the extreme contrast between the leafy Kya Sands suburb and its neighbouring informal settlement, shown in Figure 1.2.

Latin America (which includes Argentina, Brazil, and Mexico), also has a poor Gini Index according to Figure 1.1. A decade ago their ratio was even worse. This trend has, however, since been turned around. In 2011, research was conducted to determine the cause of this decline in inequality [73]. One of the two main factors identified was the decline of the so-called *skill premium* in these countries. Skill premium refers to the difference between the wages earned by skilled and unskilled labour. One of the main reasons for a decline in the skill premium of a country is an increase in its skilled labour. This increase leads to an oversupply of skilled labourers which subsequently leads to a decreased supply of unskilled labour. This may lead to a higher demand for unskilled labour, which subsequently results in wage increases within this sector. This study suggest that a decline in inequality is a result of increase in skilled labour, and in order for a country to increase its skilled labour, a larger portion of the population has to pass successfully through the tertiary education system. The *United Nations Educational, Scientific and Cultural Organisation* (UNESCO) recently echoed this statement by declaring that a lack of qualified labour is one of the most important factors constraining growth and the possibility for sustainable development within a country [234]. The quality of local tertiary education may therefore be seen as critically important for the development of South Africa.

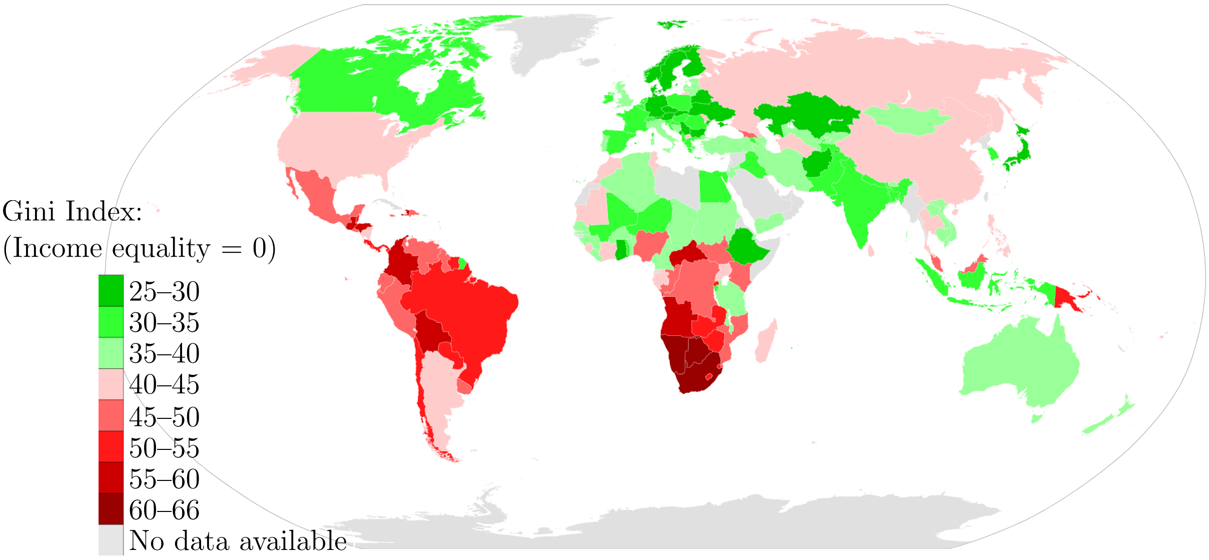


Figure 1.1: *The Gini Index of income inequality distribution in 2014 per country according to data obtained from the World Bank [224].*

While the quality of tertiary education is, of course, important in this sense, actually gaining access to the tertiary education system by having financial barriers removed is equally important and is becoming increasingly more so in view of the 2015 *Fees Must Fall* campaign [158]. A variety of *non-governmental organisations* (NGOs), governmental organisations and private corporations assist underprivileged individuals in this respect by channelling funding for their tertiary studies to these individuals. A Cape Town-based NGO which provides such financial support specifically to individuals originating from poor, rural communities within South Africa is the main industry partner of this study. Its sponsors, who provide financial donations for bursary purposes, include a wide variety of both private and governmental bodies.

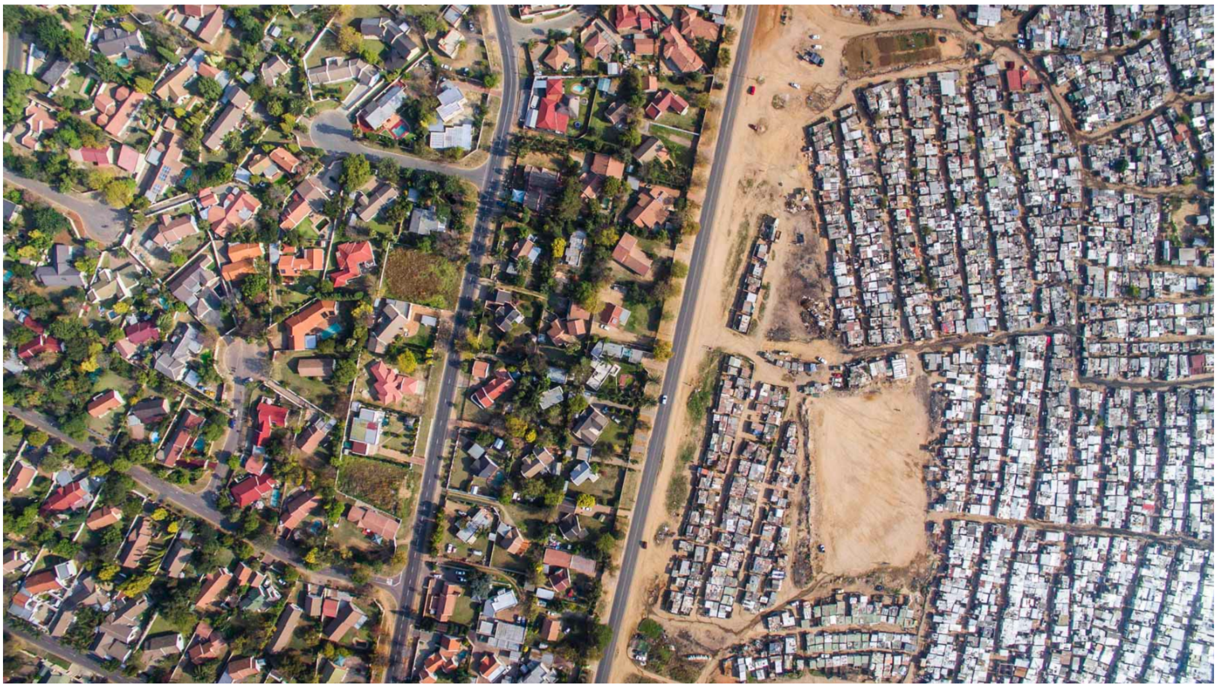


Figure 1.2: *An aerial view of unequal living conditions in two suburbs in Kya Sands, Johannesburg, South Africa [148].*

The NGO, henceforth referred to as the *main industry partner*, currently uses a very basic method of weighted criteria when selecting candidates earmarked for financial support in the form of bursaries. Its current application process requires that recruiters go into rural areas where they assist candidates with their bursary applications. These recruiters collect information about the applicants which forms the basis for three assessment scores. These scores, along with applicants' school marks, are the main factors currently considered by the NGO's selectors when selecting applicants for financial support. Although the current process works reasonably well, it is not optimised and a need for streamlining the process has been identified. An appropriate improvement methodology resides within the realm of statistical models, machine learning, and *multi-criteria decision analysis* (MCDA) techniques. These techniques may form the basis of a *decision support system* (DSS), a concept demonstration of which is developed in this study.

The research reported in this thesis is, however, neither exclusively applicable to the situation of the main industry partner nor overly tailored to its specific needs. Indeed, the work in this thesis is readily applicable to other tertiary study funders of underprivileged individuals. A secondary industry partner is also engaged in order to demonstrate this claim. This industry partner, henceforth referred to as the *secondary industry partner*, also functions as an NGO and facilitates the distribution of bursaries on behalf of trusts, foundations, and private and corporate donors to underprivileged prospective tertiary students.

1.2 Thesis scope

This study entails two main elements. The first involves the investigation of sample data provided by industry partners described above in order to identify those variables that hold the most predictive power with respect to predicting the tertiary success of bursary applicants.

The second element comprises the design and development of a DSS concept demonstrator which is able to provide decision support to a user who has to select, based on specific variables, a subset of suitable candidates from a large initial set of possible bursary applicants. In the context of the industry partners of this study, this concept demonstrator is capable of providing support to their management during the process of allocating bursaries to applicants. The system is designed to allow the user to evaluate each new bursary applicant according to a set of pre-selected variables and to provide the user with the possibility of analysing certain 'what if' scenarios by including or excluding subsets of variables or statistical prediction models from the DSS computations. The aim of the DSS is not to select candidates automatically from the initial list of applicants, but rather to provide recommendations in the form a predicted outcome and a rank level for each candidate, based on the variables selected by the user. In this manner, the DSS is able to assist decision makers by decreasing the number of applicants in an initial pool of consideration so as to facilitate final selection from a shortlist of candidates. The decision maker is thereby able to ascertain which groups of candidates are the best alternatives, borderline alternatives, and high-risk alternatives. The study is limited in scope to the development, creation, testing, validation, and demonstration of a concept demonstrator of the proposed DSS which utilises sample data. The development process and DSS concept demonstrator are adequately documented so as to allow an outsourced programming company to establish and integrate a final implementation of the DSS with the servers and database of an NGO, if the NGO were to desire this.

1.3 Thesis aim & methodology

The aim in this study is twofold: First, to analyse past data for the most powerful tertiary success indicators, and secondly, to provide decision support to the management of an NGO by predicting tertiary success outcomes and providing rank levels for bursary applicants. These two aims are realised in the following manner.

The first aim is achieved by analysing past data of previous successful bursary applicants of two industry partners to determine specific indicators of potential success that best distinguish those applicants who have achieved success in their tertiary studies from those who have not. Statistical learning methods to be used in this analysis of the data include *logistic regression*, *random forest variable importance scores*, *classification and regression trees* (CART), and *frequency bar plots*. Anticipated examples of indicators of success may be applicants' Grade 12 Mathematics marks or NGO application interview scores. Other, less traditional, factors which may also prove to be able to predict tertiary success include the applicants' household incomes, tertiary accommodation set-ups or their interest in the fields of the desired qualification.

The second aim is achieved through the computer implementation of a DSS concept demonstrator founded on multiple statistical and machine learning algorithms, such as those mentioned above, as well as an MCDA outranking method. This DSS allows the user to select variables from past data according to which the DSS may teach itself how to 'think' and 'predict.' Once taught, the DSS produces both a predicted outcome of success and a rank level for new applicants, based on the same pre-selected variables. A demonstration of the working of the DSS is presented to the industry partners and their feedback on the usefulness of the DSS is obtained. This potentially also affords industry partners the opportunity of first evaluating the usefulness of the DSS before proceeding with its full and costly implementation and its integration with their systems through outsourcing. The recommendations of the DSS may also be interpreted by an NGO's management, who may use it as a basis for making more informed and objective decisions when allocating bursaries to students.

1.4 How this study relates to previous studies

A number of studies have been conducted to determine suitable characteristics for identifying certain students as potentially being more successful in academic environments than their peers. In this section, a brief review of such previous studies is presented. These studies are, however, based purely on data of the students' tertiary studies at specific tertiary institutions [7], [35], [172]. Other studies have been carried out which attempt to predict students' academic success at high school using other high school predictors [37]. In both these cases, the studies allow for a sound analysis, because the students are evaluated on the same academic platform. Neither of these types of studies is, however, applicable in the context of this thesis since the industry partners of this study are interested in ascertaining which applicants to fund *before* they start their tertiary studies.

1.4.1 Studies which have focussed on personality traits

Within the field of psychology, the characteristics typically considered in studies of the academic performance of students include students' time management abilities, self-discipline, IQs, cognitive abilities, emotional intelligence, motivational orientation, self-regulated learning skills, self-efficacy, and approaches to learning [59], [60], [70], [129], [172], [178].

Along with the above-mentioned factors, the so-called *big five* personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism are well known among psychologists in assessing the academic abilities of individuals [163].

1.4.2 Studies which have focussed on prior academic achievements

Numerous studies of tertiary success have also been carried out considering the prior academic achievements of students before they enter tertiary education as possible indicators of success at tertiary level.

The main prior academic achievement factors considered in such studies include the results of entrance examinations of tertiary institutions, national benchmark examinations, English language proficiency tests, high school subject examinations, and high school placement tests [22], [167], [214].

1.4.3 Studies which have focussed on other factors

Examples of other factors which have been considered in studies related to tertiary study success prediction include the students' ages, genders, ethnicities, disabilities, degrees of urbanisation, types of studies pursued (part-time or full-time), types of employment held by parents, parents' parenting style, socio-economic status of students' homes, types of schools attended, matriculation pass rates of schools attended, locations of schools attended, and intended degrees of study [114], [167], [213], [214], [237].

1.4.4 The nature of this study

Birch and Miller [22, p. 3] found that no single theoretical model may be used as a general template for indicating the expected tertiary success of students. This is not only due to samples of students being so vastly different between the studies cited in §1.4.1–1.4.3, but also because each of the studies seem to have been dictated to some degree by the data that were available to the respective researchers at the time.

In a similar manner the author of this thesis acknowledges that many, if not all, of the factors mentioned in §1.4.1–1.4.3 may well be significant indicators of tertiary success, and have also been proven to be so in the above studies. Regrettably, however, as Birch and Miller [22, p. 3] discovered, data on all of these factors often cannot be considered for various reasons and hence this study is also largely dictated by the available data. Four main reasons exist as to why data on all of the above-mentioned factors cannot be considered, namely that the data were never collected, that the data cannot be collected, that the data cannot be accessed, or that the data are not applicable.

Many of the research studies cited in §1.4.1–1.4.3 were conducted in countries that are predominately more developed than South Africa, which have access to the resources of adequate finances and time. Data of the quality required for these studies would necessitate academic institutions, such as schools, to have adequate infrastructure in place to facilitate consistent collection of the appropriate data. Because South Africa is a developing country, these resources are not readily available to collect the required data. Rural schools in South Africa do not, in fact, regularly collect any data of any quality, beyond regular school marks, which may assist in predicting later tertiary success. Unfortunately, much of the data required to carry out the kind of studies cited in §1.4.1–1.4.3 simply do not exist as the underlying information has never been recorded by schools.

The industry partners of this study furthermore do not have unlimited time and finances available to collect data on all of the above-mentioned factors. It would be a major expense to conduct intensive and lengthy interview processes, take down psychometric assessment results or host weekend camps with top applicants so as to assess their skills in social interaction with their peers or their working abilities in a group.

Although data related to the results of entrance examinations at tertiary institutions have proven useful in other studies, it is highly unlikely that such data can be gathered from the various respective tertiary institutions by an independent NGO, such as the industry partners of this study [22, p. 4], because such data are usually confidential.

It would furthermore be unreasonable to evaluate rural applicants based on certain factors which are influenced by their circumstances and upbringing, which are generally acceptable criteria for the allocation of bursaries in more developed countries. Although not specifically mentioned above, an example of such a factor is a rural applicant's leadership abilities and participation in sport. Information of this kind may be unreasonable for use in the context of underprivileged bursary applicants since such applicants might not have had adequate opportunities to participate and excel in leadership and sporting activities at school level.

It should finally be kept in mind that although the results of the studies cited in §1.4.1–1.4.3 may, to some degree, be informative, it will not necessarily be significant in the context of this study as those studies were conducted based on samples of students that are vastly different from those considered in this study. It would, for example, be incorrect to assume that the significance shown by the Grade 12 Mathematics marks of students from upper class schools in Australia as predictors of tertiary success at Australian universities would be similar to that of the Grade 12 Mathematics marks of underprivileged students in South Africa who attend South African tertiary institutions.

The above-mentioned challenges make this study unique in that its aim is to identify success indicators for samples specifically representing underprivileged South African high school graduates *before* they have the opportunity to commence their tertiary studies.

1.5 Thesis objectives

In order to achieve the aims set out in §1.3 within the scope defined in §1.2, the following seven objectives are pursued in this thesis:

I To *document* the literature related to this study. More specifically,

  i To *conduct* a comprehensive literature review on general *data preparation* and *information systems* (ISs) development methodologies.

  ii To *conduct* a comprehensive literature review on *statistical learning* and *machine learning* methodologies.

  iii To *conduct* a comprehensive literature review on MCDA modelling techniques and solution methodologies.

II To *collect* data and other requirements from industry partners. More specifically,

  i To *obtain* an adequate set of sample data of past performances of bursary holders from the main industry partner for analysis purposes.

  ii To *acquire* an adequate set of sample data of past applicants from the secondary industry partner for analysis purposes.

  iii To *collect* and *understand* specific industry partner user requirements, preferences, and to *document* the processes followed by the main industry partner in respect of its application process.

III To *propose* a modular and generic DSS architecture capable of combining the predictions for alternatives produced by multiple statistical learning models, in a weighted manner, and to *present* these predictions, accompanied by a ranking of the alternatives.

IV To *develop* and *create* a DSS concept demonstrator based on the proposed generic DSS architecture of Objective III. More specifically,

  i To *interpret* and *present* the process flow activities of the main industry partner's systems, as studied in pursuit of Objective II(iii), by *visualising* it by means of a *use case* and *activity diagram*.

  ii To *build* the concept demonstrator of the proposed DSS architecture of Objective III.

  iii To *implement* suitable statistical learning and machine learning models identified during the literature review in pursuit of Objective I(ii) which each individually predicts an outcome for each alternative, in an ensembled manner.

  iv To *include* an element in the DSS that assesses the validity of underlying statistical and data assumptions of the past data and makes these assessments available to the user for further consideration.

  v To *incorporate* an integration strategy, which combines the predictions of the models of Objective II(iii) into a single prediction for each alternative.

  vi To *integrate* an MCDA outranking method into the DSS concept demonstrator of Objective IV(ii) in order to produce a rank level for each alternative.

  vii To *create* a *graphical user interface* (GUI) for the above DSS concept demonstrator of Objective IV(ii) which allows for users to efficiently compare the output of different scenarios, based on the input variables chosen by them, so as to assist in an effective allocation of bursaries to applicants.

V To *perform* case studies based on data collected in pursuit of Objectives II(i)–(ii). More specifically,

  i To *determine*, by *preparing* and *mining* the past data gathered in pursuit of Objectives II(i)–(ii) and *applying* statistical and machine learning methods, those variables which are generally good indicators of success that can be used to potentially identify applicants who have a higher chance of success in their chosen directions of study.

  ii To *demonstrate*, *assess*, and *validate* the practical application and capability of the DSS concept demonstrator by applying it to the context of Objective VI(i).

VI To *report* back results and recommendations to the industry partners based on the findings of Objectives V(i)–(ii).

VII To *suggest* ideas for future work as possible improvements to the work contained in this thesis.

1.6 Thesis organisation

This thesis is partitioned into four main parts. The first, Part I: Literature study, comprises Chapters 2–4. The focus of Chapter 2 falls on a discussion of the fundamentals of ISs and data preparation. The chapter opens in §2.1 with the presentation of the various categories within which raw data may be classified. This is followed in §2.2 by a description of the steps of data preparation. An explanation is presented in §2.3 of various systems development methodologies and modelling tools according to which an IS, and specifically a DSS, may be developed. In §2.4, the topics of testing, validation, implementation, quality assurance and maintenance of an IS are discussed.

The second chapter of Part I, Chapter 3, contains a review of the fundamentals of statistical learning and machine learning. In the opening section, §3.1, the various elements of statistical learning are presented. This includes discussions on the relationship between input and output variables, regression *versus* classification, and data splitting methods. In §3.2, the reasons for correlation between variables, which might not be causal, are elaborated upon, and this is followed by an argument as to why linear regression is not appropriate for predicting binary outcomes in §3.3. The five statistical and machine learning models of *logistic regression* (§3.4), *CART* (§3.6), *random forests* (§3.7), the *C4.5 algorithm* (§3.8), and *support vector machines* (SVMs) (§3.9) are introduced thereafter. Specific emphasis is placed on the evaluation of the quality of a logistic regression model (§3.5), random forest variable importance scores (§3.7.1), and identifying outliers using random forests (§3.7.2). The chapter closes in §3.10 with an examination of the various statistical and data assumptions associated with each of the five models discussed above.

In Chapter 4, the third and final chapter of Part I, basic concepts underlying ensemble learning and MCDA are reviewed. In §4.1, the fundamentals of ensemble methods and the mechanisms for constructing such methods are presented. The different integration strategies available for combining elements of an ensemble method, partitioned into static integration strategies and dynamic integration strategies, are investigated thereafter in §4.2.

Chapters 5 and 6 together constitute the second part of this thesis, Part II: Decision support system. The focus in Chapter 5 is on presenting the proposed DSS architecture. The chapter opens in §5.1 with a visual presentation of the proposed framework and a discussion on the purpose of each interlinking component of the system. The data preparation component and the model base component of the system are discussed in more depth in §5.2 and §5.3, respectively.

The second chapter of Part II, Chapter 6, is focussed on a description of how the proposed DSS framework of Chapter 5 may be implemented. The implementation of a concept demonstrator of the DSS, following the seven phases of the well-known *systems development life cycle* (SDLC), is discussed broadly in §6.1, and in more detail in the remaining sections of the chapter. More specifically, Phases 1 (identifying problems, opportunities, and objectives) and 2 (determining human information requirements) are considered in §6.2. Phase 3 (analysis of the system needs) is similarly considered in §6.3, Phase 4 (design of the system) in §6.4, Phase 5 (software development and documentation) in §6.5, and Phases 6 (system testing and maintenance) and 7 (implementation and evaluating the system) in §6.6.

The third part of the thesis, Part III: Case studies, consists of Chapters 7 and 8. The purpose of Chapter 7 is to present the data obtained from two industry partners and to discuss the aims of four case studies to be performed in respect of these data. The chapter opens in §7.1 with a discussion on the target population for this thesis. The population is identified as those bursary applicants who fall within the South African lower class or underprivileged bracket. Thereafter, the sample data provided by the main industry partner for use in the first two case studies (which focus on variable importance analysis and identifying tertiary success predictive variables) are introduced in §7.2. In the two sections that follow, the specific aims, variable selection and data preparation processes of the two case studies are described in §7.3 (Case Study A) and §7.4 (Case Study B). In a similar manner, the sample data provided by the secondary industry partner for use in a third case study (Case Study C, also focussed on variable importance analysis and identifying tertiary study success predictive variables) are introduced in §7.5. This is followed by a discussion on the aim of and the data preparation process followed in respect of the sample data in §7.6. The aim of a fourth case study, Case Study D, which focusses on assessing the performance of the DSS developed in Chapter 6, is disclosed in §7.7.

The four case studies outlined in Chapter 7 are performed in Chapter 8. Case Study A, concerned with answering a specific question pertaining to the importance of three score variables currently used in the bursary application process of the main industry partner is conducted in §8.1. A nine-step roadmap for investigating interesting trends in the data and identifying those independent variables in respect of which tertiary studies success prediction may be attempted is outlined in §8.2. Case Studies B (§8.3) and C (§8.4) are performed according to the methodology laid out in this roadmap. The fourth and final case study, Case Study D, is performed in §8.5.

The final part of the thesis, Part IV: Conclusion, comprises Chapters 9–10. Chapter 9 contains a summary of the work presented in this thesis (in §9.1) and an appraisal of the contributions of the thesis (in §9.2). The thesis closes with a number of ideas for future work in Chapter 10.

Part I

Literature study

CHAPTER 2

Information systems and data preparation

Contents

2.1 Raw data categories 13
2.2 Data preparation 15
2.3 Information systems 18
2.4 System quality, testing, implementation and maintenance 27
2.5 Chapter summary 30

The goal of this chapter is to review the literature on the tools, techniques, approaches, and methodologies required to design, evaluate, and implement an information system. First, the concept of data and information is discussed. Thereafter, the different categories to which data may belong is presented. This is followed by a discussion on how data may be prepared for use in a study.

Next, information systems are considered, with a specific focus on decision support systems. The main components of a decision support systems, the database, the user interface, and the model base are described in some detail. This is followed by an explanation of the systems development methodologies and modelling tools that may be used to develop an information system. Three such systems development methodologies are described in more detail, namely the waterfall methodology, the agile methodology, and the object-orientated methodology. The final topic in the review of this chapter centres around how to maintain the quality of, test, validate, implement and maintain an information system.

2.1 Raw data categories

According to Baltzan and Phillips [15, p. 40], information may be seen as raw data that have been converted into meaningful and understandable content which may then be used to support the business operations of an organisation.

Raw data may be grouped into two main categories according to the properties of the underlying variables, namely *quantitative* data (Figure 2.1 (a)), and *qualitative* or *categorical* data (Figure 2.1 (f)). As this study involves the analysis of both quantitative and qualitative data, it may be referred to as a *mixed methods research* study [58]. Curry *et al.* [47, p. 1442] and James *et al.* [99, p. 6] describe quantitative data as data that are located on a numerical scale, such as the speed of a car or the academic average of a student. While all quantitative data are numeric, not all numeric data are quantitative. For example, the national identification number of an individual is numeric, but not quantitative. Qualitative data, on the other hand, may be classified into categories based on the inherent characteristics of the objects described by the data, such as a five-star hotel rating system or the gender of a person. Qualitative and quantitative data may further be divided into the categories shown in Figure 2.1. Each of these categories is described in some detail in this section.

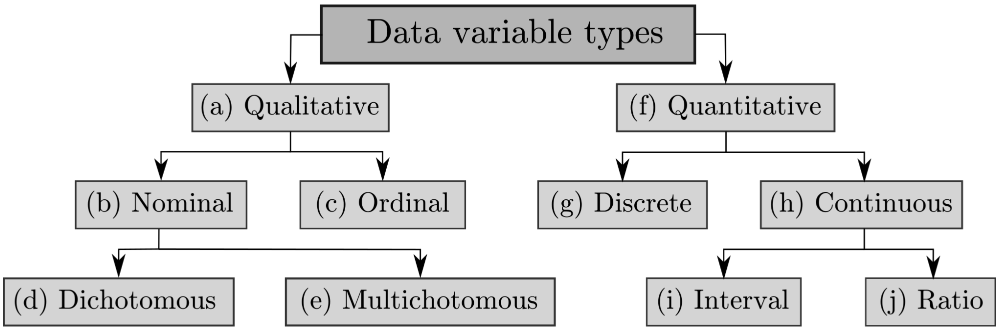


FIGURE 2.1: *Data variable types.*

Two types of qualitative data exist, namely *nominal* data and *ordinal* data. Hastie *et al.* [88, p. 504] explain that nominal data (Figure 2.1 (b)) contain two or more categories which may or may not be arranged in a meaningful sequence, but which cannot be quantified or ranked. Nominal data which are in an 'either-or' format and only have two categories are referred to as *binary* or *dichotomous* data (Figure 2.1 (d)). An example of such data is the gender of a person: either male or female. Any other nominal data, with the same properties, but which occur in more than two categories is referred to as *multichotomous* data (Figure 2.1 (e)). For both dichotomous and multichotomous data no single category may be seen as being more important than any other, nor may it be said that these categories have an intrinsic order [128].

According to Bramer [27, p. 12], ordinal data (Figure 2.1 (c)) is the second kind of qualitative data. This data type is similar to nominal data in the sense that these data occur in two or more categories which cannot be quantified, but unlike for nominal data, the categories of ordinal data are intrinsically ranked. An example of this type of data may be found in a survey where people were required to indicate how often they play a specific sport: "Never," "Not very often," or "Regularly." Each one of these possible results may have a numerical value associated with it (*e.g.* 0, 1 and 2). It should, however, be noted that although an intrinsic order exists over the categories, it may not be stated that the difference between the 0 and 1 categories is the same as the difference between the 1 and 2 categories [88, p. 504].

There are also two types of quantitative data. The first, *discrete* data (Figure 2.1 (g)), can assume any of a countable number of values or values which are isolated and separated by gaps (such as the number that may be rolled on a die: 1, 2, 3, 4, 5 or 6). The second type, *continuous* variables (Figure 2.1 (h)), is measurable along a continuum and may therefore take on an uncountable number of values (such as the temperature of a specific location). Any quantitative variable which cannot assume an uncountable number of values between the maximum and minimum of a specified range necessarily gives rise to discrete data [153].

Continuous data may further be classified as either *interval* data (Figure 2.1 (i)) or *ratio* data (Figure 2.1 (j)). According to Baroni and Evert [16, p. 17] and Lund and Lund [128], ratio data have all the properties of continuous data, but must also have a true zero. According to Bramer [27, p. 13], this true zero should reflect the absence of the specific measured characteristic (such as the height of an individual in centimetres which, if zero, implies that the individual has no height). Interval data may be considered as continuous data, but without the property of having a true zero (such as temperature measured in degrees Fahrenheit (°F) or degrees Celsius (°C), as zero °C or °F does not imply no temperature). It is, however, possible for temperature to be expressed as a ratio variable if measured in Kelvin, since zero Kelvin indicates that there exists no temperature.

Another important difference exists between interval and ratio data [198]. Interval variables remain meaningful if they are added or subtracted from one another, but not if they are multiplied or divided by each other. For example, the difference between 20°C and 30°C is the same as the difference between 60°C and 70°C, but it is not true that 80°C is twice as hot as 40°C. Since ratio data have an absolute zero they remain meaningful if altered by addition, subtraction, multiplication or division. For example, if the mass of an object is considered, the difference between 10kg and 20kg is the same as the difference between 80kg and 90kg. In addition, an 80kg object is twice as heavy as a 40kg object.

2.2 Data preparation

This section contains a discussion on the concepts associated with data preparation, which includes the processes of data gathering, data integration, data extraction and data cleaning.

2.2.1 Data gathering

Before any data analysis may begin, the relevant data have to be gathered from all the various sources and industry partners taking part in the study. In theory, these industry partners and sources will provide the analyst with complete *data sets*. Such data sets may typically be visualised in a tabular format where the columns represent *variables* or *attributes*, and where the rows represent *observations*, *instances* or *records* [27, p. 12].

Unfortunately, companies often do not have well-developed and carefully maintained information systems to keep their data organised. The first, and often largest, challenge associated with data preparation is to actually obtain the data relevant to a particular study. The data sets received are typically unorganised and inappropriately structured. Furthermore, these data sets are usually ridden with redundant entries and may be incomplete, which may cause confusion when analysing these data sets [160, p. 51].

2.2.2 Data integration

According to Nisbet *et al.* [160, p. 39], the second challenge in data preparation is normally to integrate data sets from various sources into a single set. These various sources may belong to the same company, but the various data sets may be located in different databases within different departments. This is often a significant challenge as data from different sources typically come in different formats, exist in different levels of aggregation and may be expressed in different units. The goal is to adapt the format of the various sets of data in such a way that they may be presented in a unified format and structure.

2.2.3 Data extraction

Once all the relevant data have been obtained and integrated into a unified format, the next challenge in data preparation is to customise the data structure in order for it to be used as input to data analysis software [160, p. 39].

2.2.4 Assessing data quality and performing data cleaning

The final challenge in data preparation is to assess the data quality and to perform the necessary data cleaning. Although data may have been converted into formats that may be used as input to appropriate data analysis software, Bramer [27, p. 15] states that it would be foolish to assume that the entire data set is without errors. Hellerstein [90, p. 1] concurs and adds that the issue of poor data quality remains a problem for many organisations. According to Nisbet *et al.* [160, p. 40], data cleaning entails the three main focus areas of *imputation* (*i.e.* filling blank entries) handling error values, and treating outliers. These three areas are discussed in some detail in the remainder of this section.

Missing values

Since many statistical algorithms which focus on prediction or classification can only use data records that contain entries or values for all the attribute columns [160, p. 40], it is necessary to address the problem of missing values. Bramer [27, p. 17] provides two methods for dealing with missing values in data sets. The first method is easier to execute than the second and entails the deletion of the record or row of data in question. Only the remaining, complete entries in the data set are often used for analysis. This strategy ensures that no erroneous data appear in the clean data set and is the logical approach when the proportion of entries that contain missing data is small. As this proportion becomes large, however, employing this method of data record deletion is poor practice, since the integrity of the data may be compromised when many records are excluded from the data analysis.

The second strategy entails replacing a missing data value with the most commonly occurring or average value of the attribute in question. This method should, however, be approached with much care. In the case of quantitative variables, the average value of the remaining entries is typically used, while for qualitative variables the most commonly occurring instance among the remaining variables is used. If the spread amongst the occurrences of the possible options for a quantitative variable is unbalanced, this method is more easily justifiable, as opposed to when the spread is reasonably balanced. An example of a reasonably balanced spread would be a variable that may take one of three possible values and the percentage of occurrences of each category is 35%, 35% and 40%.

Other methods, such as the *association rule*, *maximum likelihood imputation*, and *multiple imputation*, also exist for replacing missing values [160, p. 60].

Error values

Bramer [27, p. 15] states that data entries which contain errors may be partitioned into two sets, namely those which are *invalid* and those which are *valid*. Valid data errors refer to cases where the values entered are not the correct values, but where the system will not necessarily flag it as incorrect. Suppose, for example, that a weight of 957 kilograms is entered for an individual who is part of an athletics sports team. Unless the database contains a filter for detecting entries which are larger than a suitable threshold, the occurrence will not be flagged as an error since the only requirement for the field is that it must be numerical.

Invalid data errors, on the other hand, refer to the situation where a variable value entered does not conform with the characteristics and specifications of the variable field. Consider, for example, the value “56.8k” entered for a particular adult’s weight. In this case the “k” should not be present since the relevant field should only contain numerical values. If errors of this kind occur, they should be corrected. In this case the “k” should be removed from the “56.8k” entry. If the error is not obvious or clear enough (such as when a data entry “160” is entered for a quantitative field describing the percentage mark obtained by a student for a subject), the entry should be deleted [160, p. 57].

According to Hellerstein [90, p. 2], error values are the result of one of four main activities which should be considered when analysing data. The first activity is errors occurring from *data entry*, or human activities, since manual labour is typically employed in data capturing processes in industry. This activity includes typographic errors and misinterpretation of raw data. The second activity is *measurement* and refers to cases where data measuring instruments are used incorrectly or the entire sampling and measuring strategy is approached incorrectly. The third activity is *distillation* and occurs when errors are made due to data sets being preprocessed and summarised before they are added to the final database. This often occurs when database managers want to simplify or reduce the volume of data. The fourth and final activity is *data integration* (already mentioned in §2.2.2) and refers to errors that occur when data from various sources are integrated into a single database.

Outliers and influential observations

The last of the three main focus areas of data cleaning identified by Nisbet *et al.* [160, p. 40] is the handling of outliers. In addition to outliers, Little and Silal [125, p. 33] also identify influential observations as a cause for concern.

According to Moore *et al.* [155, p. 16], an outlier may be defined as "*an individual value that falls outside the overall pattern*," while Little and Silal [125, p. 33] define it as a data point that is not fitted well by a regression. If a data point's inclusion in or removal from a model has a considerable impact on the model outcome it is an *influential observation*. As noted by Parke [171, p. 153], however, an outlier is not necessarily an influential observation and *vice versa*.

Aguinis *et al.* [3, p. 287] identify three steps that should be followed during the treatment of outliers in data, namely *definition*, *identification*, and *handling* of the outliers or influential observations. Many statistical software packages have built-in functions which may be used for identifying outliers. They typically use a variety of outlier detection algorithms [160, p. 57]. These algorithms are typically based on single-construct techniques which may include a box plot analysis, a stem and leaf plot analysis, a schematic plot analysis, a standard deviation analysis, or a percentage analysis. The algorithms may also be based on multiple-construct techniques, which may include a scatter plot analysis, a q-q plot analysis, a p-p plot analysis, a standardized residual analysis, and or a studentized residual analysis [3, p. 276], [125, p. 33].

Little and Silal [125, p. 33] suggest a method for identifying influential observations by using Cook's distance. Cook's distance allows for the assessment of the influence that a single data point has on a regression by determining how much the regression changes when the data point in question is deleted. Once Cook's distance has been determined for each observation of the sample, a cut-off value must be chosen, above which an observation may be considered influential. Although other suggested cut-off values are mentioned in the literature, a generally accepted cut-off value is 1, and a more strict cut-off value is 0.33 [24], [44].

Let $\hat{Y}_j$ denote the prediction of observation j for a full regression fit, and let $\hat{Y}_{j(i)}$ denote the prediction of observation j for a regression fit which excludes observation i. Also let MSE indicate the mean squared error of the regression model and let m be the number of independent variables and n the number of observations present in the regression model. It is then possible to define Cook's distance for observation i as

$$\mathrm{CD}_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(m+1)\mathrm{MSE}}.$$

Another far simpler, but much more subjective method of finding outliers, suggested by Little and Silal [125, p. 33], is to search for observations which are relatively large compared to the others.

During an interview with the New York Times in 1990, Edward Ng famously stated that "one man's noise is another man's signal" [23]. This statement ties in with the finding of Aguinis *et al.* [3, p. 271] who mention that, although there is an agreement among authors in the literature that outliers are a serious concern when dealing with data-related issues, there currently exist no standardised guidelines on how to handle outliers in data.

There are two fundamental ways of dealing with outliers: either to keep them or to delete them. If the goal of a study is to identify rare or outlier events, such as people defaulting on their credit card payments, the outlier data are vital for the success of the study and should be kept. In other cases, the preferred option is that outliers should be removed since the goal is to build a model which predicts normal responses (in which case outliers will inject noise and may 'confuse' the model). The logical argument is that outlier data should be present in a data set so as to be able to predict similar outliers. The counter-argument is that it is fine to loose some predicting ability (*e.g.* 3%) in order to gain predictive ability on the rest of the data set (*e.g.* 97%). Either way, it should be clear that the way in which outliers are dealt with has to be considered carefully for the specific case at hand [160, p. 57].

It should be noted that if data points are deleted, the findings of a study may potentially lose credibility. The reason for this is that conclusions drawn from predictive studies with regard to how independent variables affect dependent variables may well change significantly depending on how outliers are handled. According to Aguinis *et al.* [3, p. 272], if deletion of outliers does take place, it should be noted in the footnotes of the respective publication.

2.3 Information systems

In order to manage this valuable resource of information, organisations construct *information systems* (ISs) capable of efficiently collecting, analysing, storing and sharing data. In this way, such a system can support the management in its business operations and decision making by assisting it in better understanding and controlling its business processes and operations [56, p. 1].

ISs are considered in this section, specifically focussing on the manner in which they may be utilised to store and manage data. This is followed by a discussion on how ISs may be developed.

ISs may be classified in a variety of ways, but the most commonly used classification model is the four-level pyramid model[1] illustrated in Figure 2.2. This four-level model classifies ISs based on the level of staff that will use it. By employing this classification method, both the nature of the task and the system users' abilities are indirectly considered [106].

The main types of ISs commonly employed are listed inside the four-level pyramid in Figure 2.2. These ISs range from transaction processing systems used to process large amounts of data for routine work by ordinary floor staff, to executive ISs used by executives to grasp or gain a better understanding of their environment so as to be equipped to solve unstructured decision problems [103, p. 30].

[1] Other variants of the model exists that classify the ISs slightly different. Two of these are the *three-level pyramid model* and *five-level pyramid model* [106].

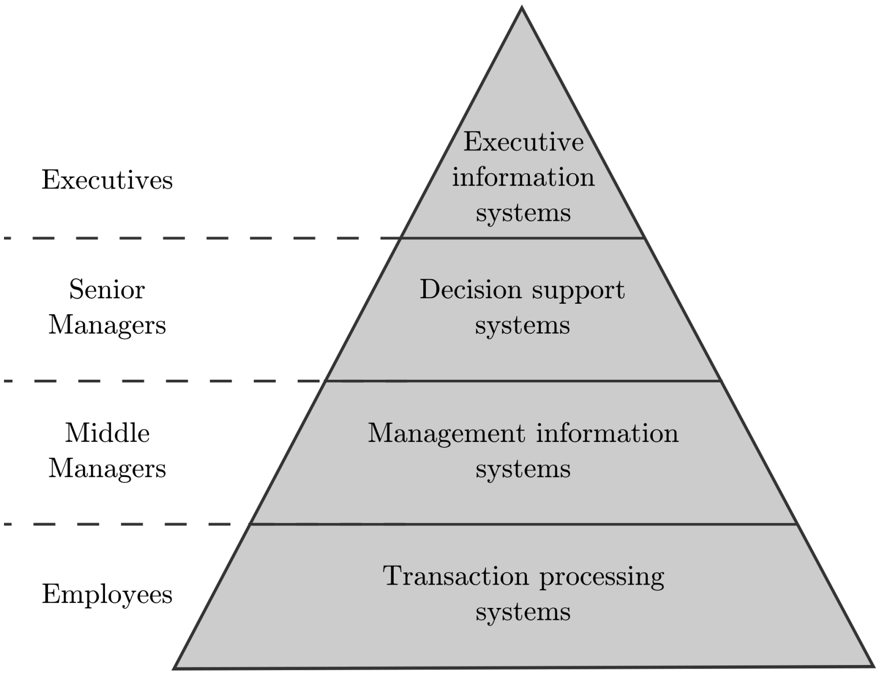

Executives
Senior Managers
Middle Managers
Employees
Executive information systems
Decision support systems
Management information systems
Transaction processing systems

Figure 2.2: *The four-level pyramid model adapted from Kimble [106].*

2.3.1 The role of a systems analyst

An individual who endeavours to generate value for his or her company, according to Dennis *et al.* [53, p. 2], may be defined as a *systems analyst*. In addition, Kendall and Kendall [103, p. 34] suggest that such an individual should have a natural inclination to solving problems and to analyse a company's systems with the intent of obtaining a greater understanding of its data flows. In addition, a systems analyst typically utilises his or her past knowledge and skills of relevant methods and tools to create a new IS or improve an existing IS. The systems analyst should work closely with the relevant parties, especially the final user, by actively involving and adequately communicating meaningful information to them during the development process.

2.3.2 Decision support systems

The IS type that falls within the second tier of the four-level pyramid model in Figure 2.2 is known as a *decision support system* (DSS). According to Doucet [56, p. 2] and Kimble [106, p. 6], the function of DSSs are to support varying sized groups of people within a company by presenting them with results that otherwise would be difficult to obtain without the aid of the system.

According to Kimble [106, p. 6], any DSS should be designed and constructed while taking the skill level of the final user into consideration, as the final system will only be of real value to the user if he or she is able to comfortably understand the system's output. This supports the earlier statement that the various types of ISs are classified in terms of their users' current skill levels.

Doucet [56, p. 2] mentions that most ISs, and thus also DSSs, consist of three essential elements. These are the *database*, the *graphical user interface*, and the *model base*.

The database

A database allows for the structured storage of related data, from where it is accessible to the other relevant segments of a DSS [103, Chapter 13], [242, p. 6]. A well-designed database is crucial for the performance of a DSS [142]. Four of the benefits associated with such a design, according to Watt and Eng [242, p. 10], include the following:

- No data redundancy (ensuring that no duplicates are created in the data),
- Concurrency (*i.e.* multiple users are able to access the data at the same time),
- Data integrity (ensuring that no incorrectly formatted entries enter the database), and
- Security (so that it is easy to manage data access restrictions).

The manner in which data are stored and managed in a database depends on the type of database model chosen for use by the systems analyst. According to Obbayi [161], five main types of database models exist, namely *flat file-based* databases, *relational* databases, *hierarchical* databases, *network* databases, and *object-orientated* databases.

Possibly the most uncomplicated of the five types of databases are flat file-based database models. The data contained within these databases are in binary or *human-readable* text formats. They are popular for stand-alone applications, but require that the records or observations within the data all have entries, *i.e.* no missing values allowed. Such a database model often makes use of `Microsoft (MS) Excel` files, or more specifically *comma separated value* (CSV) file types.

The most popular database model, however, is the relational database model. This model approach involves *normalising* data and storing the results in various tables. Such a model is designed using *entity-relationship diagrams* (ERDs) in which all systems are considered as entities and relationships, and which make use of keys to logically link the different tables together. The relationships between the entities can be either one-to-one, one-to-many, or many-to-many. A popular way in which ERDs are created and altered is through the use of the *structured query language* (SQL). Although less efficient than other models, with adequate processing power it is generally not an obstacle [33, p. 279], [41, p. 16], [103, p. 435].

Hierarchical database models, are tree-like models which operate in a parent-child manner with the relationships typically being of the kind *one-to-many*. This database model works well for records with many describing attributes and which can be stored in a nested format. A drawback of this database model is that it constantly needs to adjust and rebalance as changes are made, resulting in low efficiency [161]. A popular method to store data according to this model is by use of the *extensible markup language* (XML) formatted files.

The network database model is similar to the hierarchical database model except that the relationships between the parents and children may also be of the kind *many-to-one*. This database model type generally uses SQL for data manipulation and was popular during the 1960s and 1970s, but has since mostly given way to relational database models [161]. As this model forces a search for a single item to look through the entire data set it is also not very efficient.

The last of the five database models, and the most recently developed, is the object-orientated database model which was inspired by and is generally created with the intention of being linked with object-oriented programming languages. According to this model, data are not stored in relational tables, but are rather partitioned into objects from where the data are accessed using *pointers* (locations within the data of the entries in question).

With the five above-mentioned main options available for a database model it should be clear, as noted by French [69, p. 449], that it would be incorrect to assume that there exists a one-size-fits-all type of database. It is rather the case that each database model has its own strengths and weaknesses, and that an appropriate model should be selected and possibly tailored to suit the specific need of the organisation.

The graphical user interface

Human-computer interaction (HCI) is made possible by what is known as a *user interface* (UI) or *graphical user interface* (GUI). It provides a link between the human operator and the computerised decision supports, and allows users to interact with the system. The user typically provides all the inputs required by the system and obtains all the relevant outputs via this GUI.

According to Kendall and Kendall [103] and Stone *et al.* [216], the main quality required for a GUI to be considered 'good' is *usability.* A usable GUI is one that the user may easily engage with and one that grants the user the ability to effectively and efficiently perform his or her required functions.

In addition to usability, Benyon [18] mentions that it is necessary to strive for harmony between overwhelming users with information and possible choices, *versus* discouraging them with an unclear GUI. Thus, a GUI which withholds unnecessary intricacies will most likely provide greater contentment on the part of the final user [225]. Although there exists truth in the above statements, Mandel [133] issues a warning that, despite the above, the systems analyst should create a GUI by forfeiting potentially quicker ways to perform certain procedures purely for the sake of not making the GUI overly complex, as the additional steps of an overly simplified GUI will cause frustration for more competent users. Once again this draws attention to the need to understand the users' capabilities and the importance of continuously consulting with the final users during the development of the GUI so as to obtain harmony between these two extremes.

When considering the design of the interface, the user-centred design illustrated in Figure 2.3 (adapted from Stone *et al.* [216]) may be used to involve the final users in a continual fashion. Mandel [133] agrees by stating that through a continual relationship it will be able to create a GUI that is custom-built for the final user, as opposed to an interface to which users are required to adapt. Mandel [133] also identifies three rules that should be followed during the design of a GUI. The rules are: give the user adequate authority over GUI, decrease the user's cognitive load, and create a consistent GUI so as to avoid additional confusion.

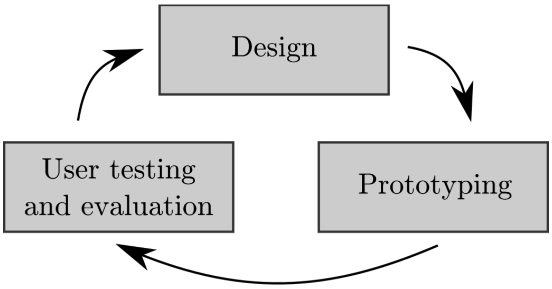


FIGURE 2.3: *An iterative UI design and evaluation process, adapted from Stone et al. [216].*

Model base

The model base component contains the models that are used to solve problems and present different alternatives during the decision making process. Such models may appear in many different forms, such as an individual mathematical algorithm or combinations of various algorithms, techniques, or methods. Such a combination gives rise to the problem of how to combine the output from the various models and/or techniques. Malhotra *et al.* [132, p. 2] cite many studies in which more than one model or technique was used to analyse data, but mention that the various models were analysed in a competing manner. This entails presenting a variety of models with the same data, evaluating the strengths and weaknesses of all the models, but only accepting the output from one model for all of the observations.

Malhotra *et al.* [132, p. 2], on the other hand, suggest a framework for combining the output from various models and/or techniques in a complementary rather than a competitive manner. The framework is presented graphically in Figure 2.4 and was designed specifically for the classification of observations. The solid lines represent data flow during the process of classifying the observations and the dotted lines represent data flow of the observations on which a consensus was not initially reached.

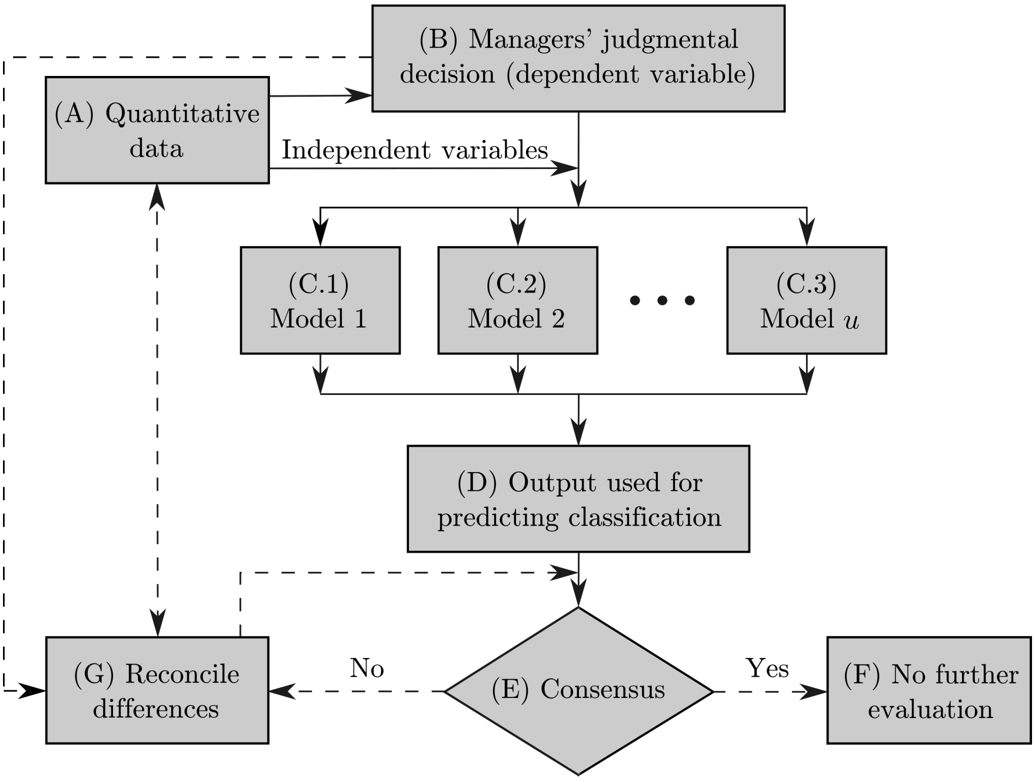

(B) Managers' judgmental decision (dependent variable)
(A) Quantitative data
Independent variables
(C.1) Model 1
(C.2) Model 2
(C.3) Model u
(D) Output used for predicting classification
(G) Reconcile differences
No
(E) Consensus
Yes
(F) No further evaluation

FIGURE 2.4: *Framework for decision making using multiple models, as presented by Malhotra et al. [132, p. 3].*

Within the framework, the process starts with the quantitative data (A) from which the possible dependent variables are extracted by the relevant managers or subject experts (B) who evaluate them so as to select appropriate output or dependent variable categories. With the dependent variables decided upon, the data set, along with the independent variables, are passed on to all of u models (C.1–C.3).

Each of the u models will then produce a predicted output (D) for each of the observations within the data set. For each observation's output in the data set it is investigated whether there is a consensus among the models (E). The evaluation is complete for the observations for which there was a consensus (F). The observations for which a consensus was not reached, however, need to be re-evaluated individually (G) by the managers and subject experts so as to reconcile the differences and decide on an output for those observations. This re-evaluation may often take place in a more qualitative, as opposed to quantitative, manner. It may also be necessary to gather additional data on these cases.

This framework has the added benefit of separating the observations for which a clear consensus is reached from those which need more attention, and allows such additional attention only to be focused on the observations which require it, as opposed to the entire data set [132, p. 4].

The software

Shelly and Rosenblatt [200] argue that an IS brings information technology (which includes software), people, and data together. Therefore, software (although not one of the three main components of an IS) nevertheless forms an integral part of an IS as it allows all of the components to be integrated.

When it comes to selecting software to use for a particular project, the following six factors are important considerations:

1. *Licensing* refers to whether the software is free or has to be purchased [62].
2. *Connectivity* describes whether the DSS created will operate offline or online.
3. The *language on server side* is the computer language used to implement the algorithms. Examples of such programming languages include `ASP.NET`, `C`, `Java`, `PHP`, `Python`, `R`, and `Ruby` [203].
4. The *data store* refers to the software used to store the data, such as `MySQL` or `MS Excel` [77].
5. The *integrated development environment* (IDE) assists the programmer by reducing the configuration required during programming and thereby streamlining development. Many IDEs are designed for use in conjunction with specific programming languages [164].
6. The *web application framework* promotes code reuse by providing libraries and templates for use during programming. Web application frameworks are also designed for use in conjunction with specific programming languages [111].

Within each of these six categories the systems analyst has to make choices during the system development process. There is generally not a single correct choice for any of the software categories, but rather the selection should be made by taking into consideration the factors of available time, finances, software, user requirements, user skill level, and personal preference.

2.3.3 Systems development methodologies and approaches

In order to develop a DSS as described above, or any IS, *systems development methodologies* (SDMs) or *systems development processes* may be employed. An SDM is defined by Walters *et al.* [241] as an array of operations, tools, techniques and documentation methods that may assist systems analysts in the creation of ISs. SDMs typically share a number of core phases in the development process, known as the *systems development life cycle* (SDLC) [53, p. 3], [103, p. 36]. Although experts do not agree upon how many phases exists in the SDLC, Kendall and Kendall [103, p. 36] identify seven of these phases, as illustrated in Figure 2.5.

Although the SDLC in Figure 2.5 forms the foundation of systems development, variations may occur, because development teams typically choose to approach the designs of their particular systems in different manners [53, p. 3], [197]. The major differences in the approaches are the order in which an approach specifies the way in which the SDLC phases should be performed as well as the amount of time and the number of available resources the development team wishes to allocate to each phase.

The subsections that follow contain brief discussions of the characteristics of three popular and well-documented SDMs that have been derived from the SDLC, namely the *waterfall* methodology, the *agile* methodology, and the *object-orientated* methodology.

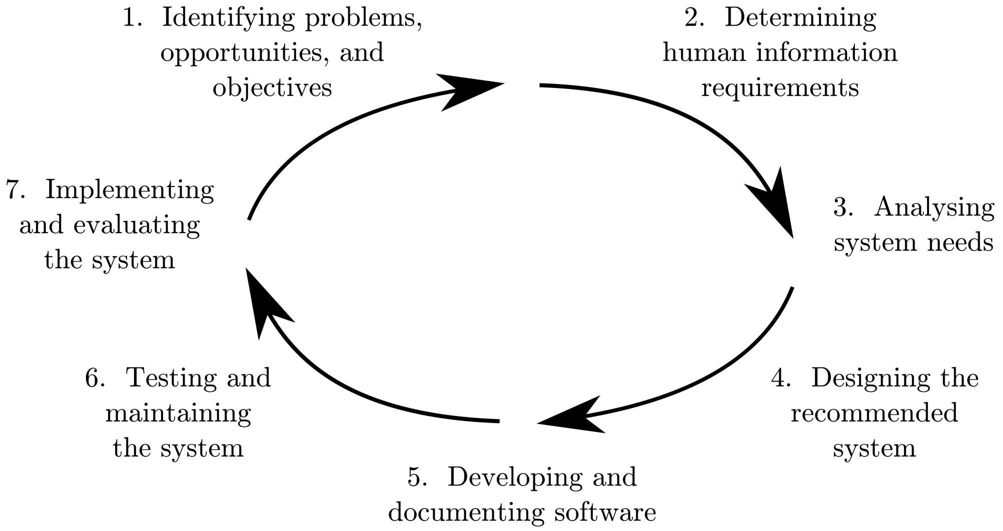

1. Identifying problems, opportunities, and objectives
2. Determining human information requirements
3. Analysing system needs
4. Designing the recommended system
5. Developing and documenting software
6. Testing and maintaining the system
7. Implementing and evaluating the system

Figure 2.5: *Seven phases of the systems development life cycle (SDLC), adapted from Kendall and Kendall [103, p. 36].*

2.3.4 The traditional structured waterfall methodology

According to Modha *et al.* [150], the last few decades have seen a number of new SDMs. The first development of such methodologies occurred during the so-called *pre-methodology era* around the time when computers were introduced to the business world. During this time, systems for computers were developed without following any defined methodologies — developers rather relied heavily on experience. In addition, these development projects were seen as short-term solutions or *ad hoc* solutions to very specific problems, rather than long-term investments to improve general decision making in organisations [13], [241].

Despite system analysts of the era having adequate programming skills, Avison and Fitzgerald [13] argue that they often fell short in respect of having true understanding of the organisations for which the systems were intended or the contexts within which the systems were to be employed. This shortcoming arose predominantly from a lack of proper communication between the system analysts and clients or final users. The lack of proper communication often produced poorly described user requirements [241] and, in turn, resulted in systems being implemented that simply did not provide the value to the user's company they could have [12].

These shortcomings left the market desiring a more structured and concise approach to the development of ISs, marking the start of the *early methodology era*, and resulted in the first of the SDMs, known as structured methodologies [12], [13]. These methodologies are identifiable by their properly documented processes and systematically arranged phases. One such structured methodology is the *traditional waterfall methodology.*

Dennis *et al.* [53, p. 8] describe the waterfall methodology as a top-down and precise approach in which the methodology's phases are completed sequentially as the entire process is properly documented. Mohammad [151, p. 3–4] adds that with the waterfall methodology it is not an option to return to a previous phase, but rather that one may only proceed forward as if descending in phases over a series of cliffs. At each of the phases the approval of the stakeholders is first required, after they have considered the progress and current documentation, before the process may proceed to the next phase — hence the name *waterfall.*

Many variations on this methodology have emerged since Royce's original model, but the version presented by Dennis *et al.* [53, p. 8] partition the waterfall methodology into four phases.

These phases are *planning*, *analysis*, *design* and *implementation*. Figure 2.6 contains a graphical illustration of this version of the waterfall methodology, adapted from Dennis *et al.* [53, p. 8].

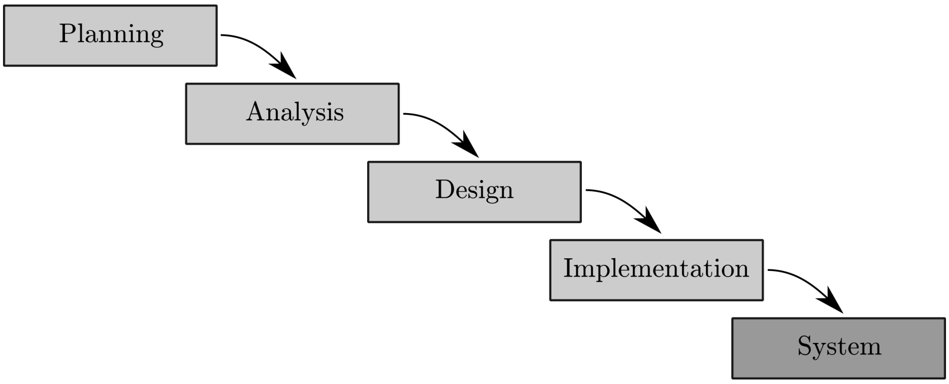


FIGURE 2.6: *The traditional waterfall design methodology, according to Dennis et al. [53, p. 8].*

Advantages of the waterfall methodology include the following. Due to the systematic development phases it is easier to manage the overall development process, as millstones may be set for each phase and measured against regular development updates. Another advantage is that due to the proper documentation the user and system requirements are known in advance and thus any programming will begin knowing the end goal [53, p. 9].

Although there is merit in these advantages, a number of drawbacks of the methodology have also been documented [53, p. 9], [241, p. 10]. It is true that benefits exist due to the priority placed on a structured development process, but these may become a liability for developers. Without much time to revise design work, the challenge of designing an entire system theoretically on paper may be very daunting. Also, if errors are discovered later in the development process, or further requirements are included, moving back up to the waterfall so as to accommodate these requirements present significant challenges. Finally, the long development process may result in a system that meets the original user requirements, but since there might have been changes to the business environment, the system is not of much use any more.

2.3.5 The agile development methodology

Following this early era, the *methodology era* emerged and with it major advances were made in respect of tools and techniques that may be utilised by systems analysts during the process of developing ISs. The additional tools included data dictionary software, project management software, and *computer assisted software engineering* (CASE) tools. These additional techniques included normalisation, entity relationship diagrams, and data flow diagrams [12].

Due to the disadvantages of the more traditional methodologies of the previous era, new methodologies emerged during the following era which made use of many of the newly developed tools and techniques [13], [241]. The first noteworthy SDM which came about during this new era was the *agile methodology*, which was created so as to speed up system development. In addition, the *agile methodology* places much emphasis on identifying and appropriately following user requirements, as well as favours systems that are practical and work as opposed to those created using much documentation [53, p. 13]. Kendall and Kendall [103, p. 42,194] argue that by using the agile methodology, faster delivery of systems is possible. In addition, developers are able to make changes and improvements to the system any time during the development, a trait that sits well with the software development world.

The major disadvantages related to the agile methodology are often more a result of incorrect implementation or human error as opposed to deficiencies of the methodology itself. Examples of these include developers demanding to take long periods of time to first document their desired plan and developers not providing working bits of code or prototypes that will demonstrate that the larger vision is indeed practical [239].

2.3.6 The object-oriented methodology

Another methodology that was forthcoming during the methodology era is the *object-orientated methodology* (OOM). According to Mohammad [151, p. 4] the OOM may be characterised as a bottom-up approach to systems development, as opposed to the waterfall methodology (described in §2.3.4) which is a top-down approach.

Structured methodologies partition the system into subsystems and in the same fashion into sub-subsystems. This is repeated to the point that no part of the system may be partitioned further. Kendall and Kendall [103, p. 45] note that the OOM partitions the system into objects or entities containing data. The data related to these objects may be locations, events, people, or actual components of the system, all of which may be grouped together into classes with similar objects that share similar characteristics.

The process described above partitions the system by employing *use case models*, which define what the system will do, without describing how it will do it and how it will interact with its environment. A *use case model* consists of four elements, namely *actors* (presented as stick men), *use cases* (presented as ovals), *relationships* (presented as lines), and the *system boundary* (presented as a box). The actors represent specific roles played by an individual or group of individuals who are related to the system. The use cases are the distinct functions preformed within the environment, whereas the relationship lines depict which actors are associated with which use cases. Finally, the system boundary represents the scope of the system in relation to the environment [53, Chapter 2], [103, p. 45].

For each use case diagram, an *activity diagram* is created, enveloping all possible use case situations. Activity diagrams, comprising swim lanes so as to illustrate which actors perform which activities, showcase the progression of, and relation between the main activities of a system. In order to display the flow of activities, various symbols are used which are placed and linked within the swim lanes. Figure 2.7 contains a representation of each symbol described above [53, Chapter 2], [103, p. 291].

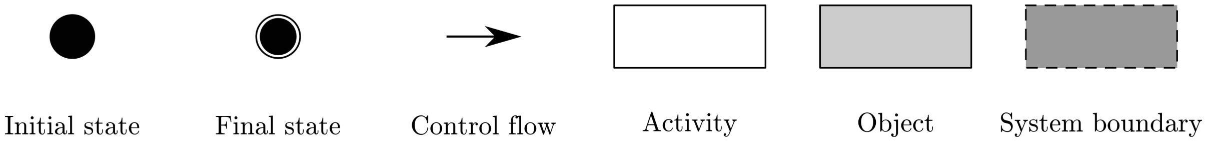


Figure 2.7: *Activity diagram elements used in the OOM.*

The two types of logical models, use case models and activity diagrams, are used in conjunction with the *unified modelling language* (UML). The UML may be defined as an assortment of tools which may be used to record the analysis phase and design phase of development, during which the various views of the system may be envisioned using specific models and diagrams. The UML is typically used in conjunction with the OOM, but may also be employed together with other methodologies [53, Chapter 2], [103, p. 45], [151, p. 4–5].

2.3.7 Criticism of systems development methodologies

The application of SDMs often provides organisations with enhanced systems and processes, but sometimes these enhancements and benefits are not achieved and thus leave certain organisations disappointed [13]. Walters *et al.* [241] and Avison and Fitzgerald [12] link this dissatisfaction to two causes. The first cause is due to the flaws of the methodologies themselves, since the selected SDM are often not properly suited to the specific companies' needs and scenarios. The second cause is that the methodologies are often not appropriately applied. This often occurs largely due to the complexity of the methodologies as the relevant knowledge about the design process and technical abilities required to apply the methodologies often do not originally exist within some organisations [12], [241].

Although the above criticism is valid in the sense that companies are sometimes not able to obtain all the befits from SDMs that they might have wished for, at the very least these companies benefit to some degree due to the use of the SDMs [12].

2.3.8 Comparison of systems development methodologies and models

There is consensus among experts in the field of ISs development that no single methodology may be considered the best — as each individual approach provides particular advantages and disadvantages [53], [103, Chapter 1], [197]. In addition, these varying strengths and weaknesses mean that each methodology will be suited differently for each potential scenario and problem. Table 2.1 is an adaptation of the SDM comparison by Dennis *et al.* [53, p. 16] and contains additional information from Kendall and Kendall [103, p. 47]. The table is a summary of how well the waterfall, agile, and OOM methodologies are expected to function in six different situations.

Table 2.1: *Aid for selecting between the waterfall, agile, and OOM methodologies.*

Ability to develop systems...	Methodology:		
	Waterfall	Agile	OOM
...with unfamiliar technology	Poor	Poor	Poor
...with unclear user requirements	Poor	Excellent	Average
...that are reliable	Good	Good	Good
...that are complex	Good	Poor	Good
...with schedule visibility	Poor	Good	Average
...within a short time frame	Poor	Excellent	Good

2.4 System quality, testing, implementation and maintenance

This section contains a discussion on how quality assurance should be applied throughout the development of a system. This is followed by a discussion on how a system should be tested, validated, implemented and maintained.

2.4.1 Quality assurance

Burnstein [34] states that there has recently been increased pressure on software developers to focus on quality issues. Kendall and Kendall [103, p. 543] provide the following three approaches that must be followed during any software development process in order to maintain quality assurance:

1. Design the system according to a top-down and modular approach,
2. Document the software using appropriate tools, and
3. Test and validate the system.

Adapting a top-down approach will assist the systems analyst not to become caught up in the minor details, but rather keep the bigger picture in mind of what the system should be able to achieve. An example of such a top-down approach, according to Ahmed [4, p. 161], is the waterfall methodology discussed in §2.3.4. According to the second point above, any software should be developed using appropriate tools. Various software considerations were discussed in §2.3.2. The third point, concerned with the testing and validation of the system, is discussed in some detail in the next section.

2.4.2 Testing/Verification

Burnstein [34] suggests that in order to develop software according to an engineering approach, one must make use of validation and verification processes, as they are vital in weighing up the quality of a system. Verification, which relates to testing, is discussed in this section, while validation is discussed in the next section.

Figure 2.8 is an adaptation from Kendall and Kendall [103, p. 554], and contains a graphical illustration of the four different types of testing that should take place during the development of an IS. In addition, the inner circle shows all the individuals who form part of the testing process (note that they are not linked to any specific type of testing phase).

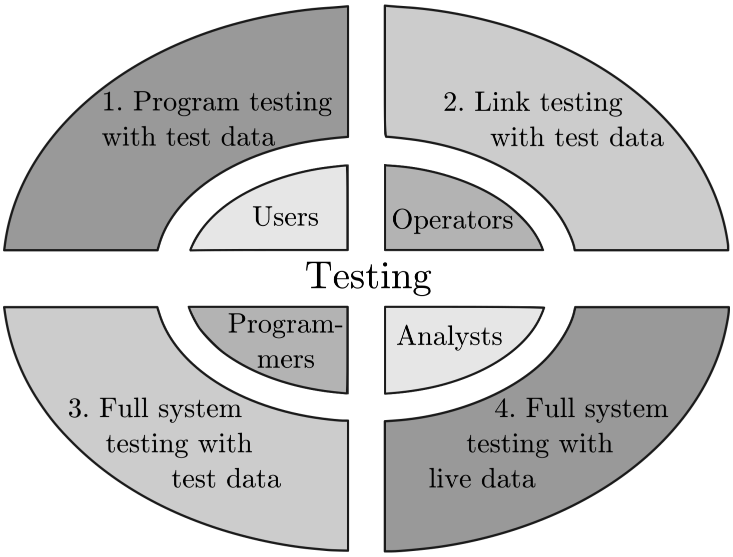

1. Program testing with test data
2. Link testing with test data
Users
Operators
Testing
Program-mers
Analysts
3. Full system testing with test data
4. Full system testing with live data

Figure 2.8: *Four steps of system testing, adapted from Kendall and Kendall [103, p. 554].*

The first step in Figure 2.8 refers to *program testing with test data*. This step entails the continuous testing that should occur throughout the building phase of a system. Every new segment of code should be tested in isolation, using small samples of test data. Test data are artificial samples of data created by the systems analyst.

The second step in Figure 2.8 is *link testing with test data.* While the previous step refers to isolated testing of small portions of code, this step refers to testing all those code portions together. Again using small samples of test data, the entire system should be tested to make sure all the segments work together as planned.

The third step in Figure 2.8 is *full system testing with test data.* In this step, the final user becomes more involved and is asked to run a few scenarios on the system, using a large set of test data. Previously, the final user was only presented with, and questioned about, individual pieces of the system. Any feedback received at this stage should be taken seriously by the systems analyst, who should make the necessary changes to the system in response to the feedback.

The final step in Figure 2.8 is *full system testing with live data.* This step requires that the system be tested again in the same manner as described in the previous step, but now with live data instead of test data. Live data refers to the use of real past data, for which the correct output is known. In this manner, the results of the new system can be compared with the results of whichever system came before it.

2.4.3 Verification and validation

Kleijnen [108] claims that there is not an agreed-upon terminology for the difference between verification and validation, but the most commonly adopted definition by Finlay and Wilson [67] is as follows. *Verification* is the process of showing that a system performs as intended, by comparing its performance against the user requirements. Verification is thus continuously assessed during the testing phase, as discussed in §2.4.2.

After verification has taken place, validation can commence. *Validation* helps to establish the client's level of confidence in a system. It is the process of ensuring that a system is a true representation of the real world, thereby ensuring that the system's output is correct and thus able to meet the real needs of the client. Both verification and validation must be performed during the development of a DSS.

2.4.4 System implementation

Kendall and Kendall [103] state that system implementation consists of two main components: first, the act of shifting computer power and responsibilities to the final users, and secondly the process of training these final users.

The process of converting to a new system may conform to *direct* changeover, *parallel* conversion, *gradual* conversion, *modular* conversion or *distribution* conversion [103].

Two parties are involved in the process of final user training — those who need to be trained and those who will train them [103]. According to Kendall and Kendall [103], these two parties should determine the strategy for training. The systems analyst has to establish four elements that are required for successful training sessions. These four elements are measurable training objectives, appropriate training methods, a suitable training site, and understandable training materials.

2.4.5 System maintenance

Gelinas *et al.* [74] and Kendall and Kendall [103] agree that the cost associated with maintenance of a system over its entire life cycle often amounts to more than 50% of its total cost. Kendall and Kendall [103] continue by stating that this financial cost is all the more reason for a systems analyst to design systems that are comprehensive and far-sighted enough to serve not only the client's immediate needs, but also those in the foreseeable future. Urhuogo *et al.* [235] add that if a system is not adequately maintained, the system and service will often be considered poor by the client. Maintenance is, however, inevitable and Gelinas *et al.* [74] mention that three types of maintenance activities exist, namely *corrective* maintenance (*i.e.* fixing errors), *perfective* maintenance (*i.e.* improving the performance of a system), and *adaptive* maintenance (*i.e.* staying up to date with business needs and responding to environmental challenges).

2.5 Chapter summary

The focus of this chapter was on data preparation and informations systems, the first part of which was dedicated to data preparation. A brief description of data fundamentals was provided and this was followed by a description of the various categories into which data may be classified. Next, the steps which should be followed when preparing data for an analysis was described.

In the next part of the chapter, focussing on ISs, the role of a systems analyst was discussed. Thereafter, a brief description of four types of ISs, and a more detailed discussion on one of those types, known as DSSs, were provided. The three main components of a DSS, namely the database, the GUI, and the model base were also briefly expanded upon. After having provided a general overview of SDMs, three well-known SDMs were described, namely the waterfall methodology, the agile methodology, and the OOM. In the final part of the chapter, the processes of assessing the quality of a system, as well as how it may be tested, validated, implemented and maintained, was discussed.

CHAPTER 3

Statistical background, learning, regression, and classification

Contents

3.1 Statistical learning 32
3.2 Reasons for correlational relationships 46
3.3 Why is linear regression not appropriate for classification? 50
3.4 Logistic regression 53
3.5 Evaluation of a logistic regression model 56
3.6 Classification and regression trees 60
3.7 Random forests 67
3.8 The C4.5 algorithm 69
3.9 Support vector machines 72
3.10 Statistical and machine learning assumptions 78
3.11 Chapter summary 86

> *"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."*
> — *H G Wells (1866–1946)*

In this chapter, various concepts from the realms of basic statistics, statistical learning, statistical regression, and statistical classification are reviewed briefly so as to provide the reader with the background necessary to understand the modelling approach adopted later in this thesis.

The first section of the chapter (§3.1) is dedicated to various general statistical and statistical learning concepts. The fundamental concept of the relationship which exists between the input and output variables of a learning model is presented first. Next, examples of such a relationship, namely the cases of simple linear regression and multiple linear regression, are presented. This is followed by a discussion on the difference between explanatory modelling and predictive modelling, as well as the difference between the two main classes of methods available for estimating output values based on input values, known as parametric methods and non-parametric methods. Next the topics of predictive accuracy of a model *versus* its interpretability, supervised statistical learning *versus* unsupervised statistical learning, and regression problems *versus* classification problems are presented. Thereafter, a distinction is drawn between the two fundamental types of research studies conducted within the paradigms of experimental research and non-experimental research. The two data splitting methods of cross-validation and bootstrapping are also discussed and illustrated by means of examples.

The next section of the chapter (§3.2) is dedicated to a discussion on the potential reasons for a correlational relationship between input and output variables. The four possible reasons for correlation are presented as: bias caused by measurement or selection error, chance relationships, confounders, and causal relationships.

The reason why linear regression is not appropriate for the purposes of classification is presented thereafter in §3.3. Five alternative classification models are presented next, namely logistic regression, classification and regression trees, random forests, the C4.5 algorithm, and support vector machines. The fundamental principles underlying each of these classification methodologies are discussed in detail in §3.4–§3.9.

The final section of the chapter (§3.10) is devoted to a discussion on the assumptions associated with regression in general, as well as those underlying the five classification method mentioned above.

3.1 Statistical learning

In this section many different fundamental concepts relating to statistical learning will be covered. First, the fundamental concept of the relationship which exists between inputs and outputs is presented, and reasons for using inputs to estimate outputs is discussed. Next, simple linear regression and multiple linear regression are expanded on. The important differentiation is then made between the concepts of explanatory modelling and predictive modelling. Thereafter, the two main methods used to estimate outputs based on inputs, parametric methods and non-parametric methods, are presented. This is followed by a discussions on: the trade-off between the predictive accuracy of a model and its interpretability, supervised *versus* unsupervised statistical learning, and the distinction between regression and classification problems. Next, the two types of research studies of experimental research and non-experimental research, as well as the difference between internal and external research, are briefly presented. Finally, two data splitting methods, namely cross-validation and bootstrapping, are discussed with examples.

3.1.1 The relationship between input and output variables

A relationship exists between the *input variables* and the *output variables* of a process [88, Chapter 2]. The input variables are also sometimes referred to as *independent variables*, *experimental variables*, *features*, *predictors*, or just *variables*, whereas the output variables are also called *dependent variables* or *responses* [57, Chapter 1]. This relationship may be expressed mathematically as

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}) + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{X}$ is a vector of input variables, $\boldsymbol{Y}$ is a vector of output variables and $\boldsymbol{\epsilon}$ is an error term independent of $\boldsymbol{X}$ with a mean of zero. As noted by Draper and Smith [57, Chapter 1], the (vector) function $\boldsymbol{f}$ is generally unknown and has to be estimated. The process of estimating $\boldsymbol{f}$ resides within the realm of *statistical learning* [99, p. 16].

3.1.2 Reasons for estimating the function $\boldsymbol{f}$ in (3.1)

James *et al.* [99, p. 17] state that two main reasons exist for attempting to estimate the function $\boldsymbol{f}$ in (3.1), namely for the purposes of *prediction* or *inference.*

Prediction refers to the process of estimating the output of a certain process with known input. If $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{f}}$ denote estimates of $\boldsymbol{Y}$ and $\boldsymbol{f}$, respectively, (3.1) reduces to $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}}(\boldsymbol{X})$ [99, p. 17]. Since $\boldsymbol{\epsilon}$ has a mean of zero in (3.1), it is eliminated when many estimates of the function $\boldsymbol{f}$ in (3.1) are made.

An accurate estimation $\hat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ depends on two factors, known as the *reducible error* and the *irreducible error*. The reducible error can be controlled, but the irreducible error cannot. It is possible, at least to some degree, to increase the accuracy by which $\hat{\boldsymbol{f}}$ is an estimate of $\boldsymbol{f}$ by selecting an appropriate statistical learning technique. The error of accuracy associated with this choice is the reducible error [99, p. 18]. Consider (3.1) again. According to James *et al.* [99, p. 19] it should be clear from this equation that $\boldsymbol{Y}$ is a function of $\boldsymbol{\epsilon}$, and thus the variability associated with $\boldsymbol{\epsilon}$ will always be present. This affects the accuracy of the estimation no matter how good it is. This effect is known as the *irreducible error*.

The second reason why $\boldsymbol{f}$ is often estimated is for inferential purposes. Within the paradigm of inference, $\boldsymbol{f}$ is also estimated, but in this case the goal is to obtain a better understanding of how $\boldsymbol{Y}$ changes as a function of $\boldsymbol{X}$ [99, p. 17]. The main goal is now not to predict $\boldsymbol{f}$, but rather to understand why and how $\boldsymbol{Y}$ changes as a function of $\boldsymbol{X}$. When predicting $\boldsymbol{f}$, the form of $\boldsymbol{f}$ is not important since the main concern is the output that is obtained from the prediction. In the case of inference, however, the shape and characteristics of $\boldsymbol{f}$ is more important, since an understanding of the working and underlying pattern of the relationship (3.1) is desired [75, p. 16], [99, p. 19]. Gelman and Hill [75, p. 16] explain that the main purpose of inference is to learn from incomplete data. James *et al.* [99, p. 19] list the following three questions that inference may assist in answering:

1. *Which predictors are associated with the response?*
2. *What is the relationship between the response and each predictor?*
3. *Can the relationship between $\boldsymbol{Y}$ and each predictor be adequately summarised using a linear equation, or is the relationship more complicated?*

3.1.3 Simple linear regression and multiple linear regression

In this section, the introduction of the previous section is elucidated by presenting two specific, popular models that are based on the fundamental concepts described earlier, namely the models of simple linear regression and multiple linear regression. Suppose, for explanatory purposes, that a sample of n observations are indexed by a set $\mathcal{N}$.

Simple linear regression

According to Alexopoulos [6, p. 23], simple linear regression involves use of a single input variable $\boldsymbol{X}$ to predict the value of a single quantitative output variable $\boldsymbol{Y}$. By adopting the linear functional form

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i, \quad i \in \mathcal{N}, \tag{3.2}$$

a single straight line will be produced as shown in Figure 3.1, where β_0 is the *intercept parameter* and β_1 the *slope parameter*. On the left hand side of (3.2), the operator $E(\cdot)$ denotes the *expected value* of its argument, and $Y_i \mid X_i$ reads Y_i *given* X_i. The expression in (3.2) therefore represents the expected value of Y_i, given a value for X_i, which may be written as a linear function $f(X_i)$ or just Y_i. As before, Y_i is the dependent variable and X_i is the independent variable [154, p. 436], [246, p. 1304].

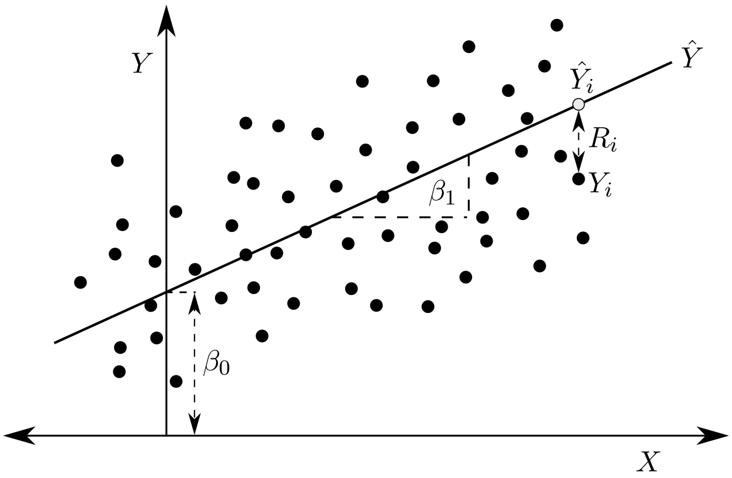


Figure 3.1: *The quantities involved in simple linear regression.*

According to Montgomery and Runger [154, p. 436], the parameters β_0 and β_1, also known as *coefficients*, fulfil the role of weights. By making use of the sample data, it is possible to estimate a relationship between the dependent and independent variable by computing estimated coefficient values. This relationship is referred to as the *estimated line* or *estimated relationship* and is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i \in \mathcal{N}, \tag{3.3}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ represent the estimated values of the parameters β_0 and β_1, respectively. The difference between the observed value of a sample datum (Y_i) and the corresponding point on the estimated line ($\hat{Y}_i$) is known as the *residual* for the pair (X_i, Y_i), that is

$$R_i = Y_i - \hat{Y}_i, \quad i \in \mathcal{N}, \tag{3.4}$$

where R_i denotes the residual of observation i, Y_i denotes the actual output, and $\hat{Y}_i$ denotes the predicted output [40, p. 30], [99, p. 62].

A popular method for estimating the coefficients β_0 and β_1 in (3.2) is the method of *ordinary least squares* [154, p. 439]. According to this method, the aim is to draw a single linear line through the data points so as best to represent them, by minimising the squared sum of the residuals.

$$SS_{\text{residuals}} = \sum_{i=1}^{n} R_i^2, \quad i \in \mathcal{N}.$$

Although the above description is a very elementary explanation of linear regression, its coefficients and residuals, it should provide the reader with an adequate notion of the basic approach [186, p. 549].

True errors *versus* residuals

The term *error* is often used to describe different types of errors in the literature. As this may cause confusion, the nature and the context of the term *error* as used within this thesis requires clarification.

The error term present in (3.1), denoted by $\boldsymbol{\epsilon}$, is known as the *true error*. This may be explained by the relationship

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i \in \mathcal{N}, \tag{3.5}$$

which is the simple linear regression relationship of (3.3) containing the true error term, also known as the *true relationship.* In theory, this relationship can be quantified using data from the entire population that is applicable to the regression line (3.5) [99, p. 63].

The difference between the *true* value found on the regression line and the *observed* value of a datum is referred to as the *true error* or ϵ_i in (3.5), which is assumed to be an independent random quantity, normally distributed, and with a mean of zero [40, p. 32]. Typically, however, only samples of the total population of data are available and since only samples are used, the true regression line and true values of β_0 and β_1 are rarely obtainable. These values are therefore sometimes referred to as *unobservable.* As explained above, however, the use of a sample of the population data may facilitate estimating a fit. The difference between residuals, denoted by R_i in (3.4), and true errors, denoted by ϵ_i in (3.5), is illustrated graphically in Figure 3.2.

It is emphasised that the error term for the population, ϵ_i, is independent of the independent variable and has a mean of zero. This means that if a large sample is used, the error term will drop away as it averages to zero. For this reason the true error term is henceforth not considered, but only the residuals [99, p. 16].

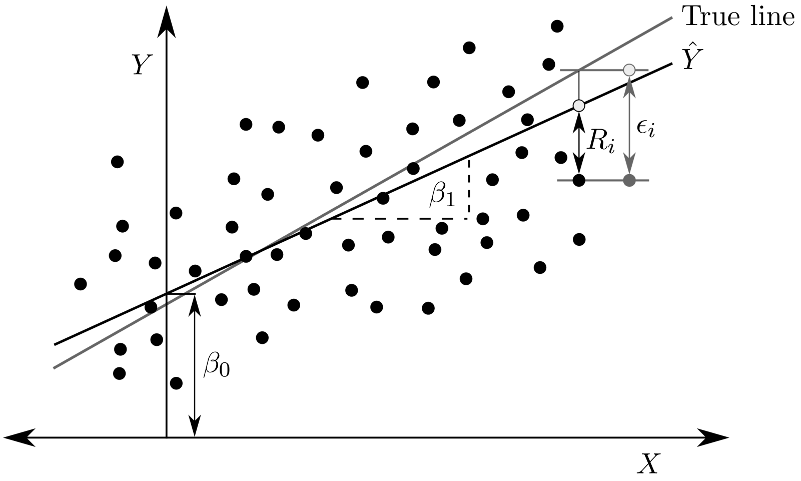


FIGURE 3.2: *Simple linear regression — residuals versus true errors.*

Correlation

If a linear relation, association, or dependence exists between the input and output of a datum, it may be said that a *correlation* exists between them. A method for measuring the strength of correlation is the *Pearson product-moment correlation coefficient*, colloquially known as *Pearson's r*, and appropriately denoted by r. This correlation coefficient assumes a value between -1 (indicating perfect negative correlation) and $+1$ (indicating perfect positive correlation). A value of 0 indicates that no correlation is present. Positive correlation implies that as the input values increase, so do the output values, while the inverse is true for negative correlation. The correlation coefficient may be calculated as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3.6}$$

for a data set $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where $\overline{X}$ and $\overline{Y}$ denote the mean of the input (or independent) variable and output (or dependent) variable, respectively [99, p. 70].

Refer to Figures 3.3–3.9 for a graphical illustration of various correlation coefficient scores for different data distributions (note that the r-scores are estimations and not exactly correct for the distributions displayed). Although a regression line with no correlation is horizontal, the slope of the regression line otherwise does not effect the score. For example, the slope of the regression lines in Figures 3.3–3.5 are identical, but due to an increasing dispersion of data points the r-value is smaller for Figure 3.4 than for Figure 3.3, and smaller for Figure 3.5 than for Figure 3.4. Also, the r-value gives no indication of the slope of the line, which is given by β_1. Indeed, two lines of different slopes may have the same r-value, as illustrated in Figures 3.7 and 3.10 [128].

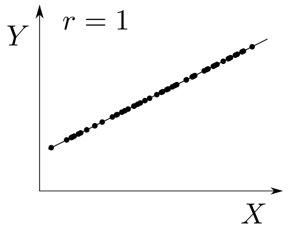


FIGURE 3.3: *Perfect positive correlation.*

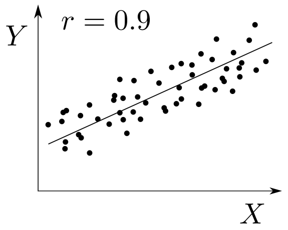


FIGURE 3.4: *High positive correlation.*

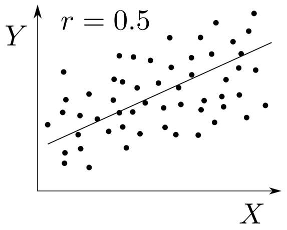


FIGURE 3.5: *Low positive correlation.*

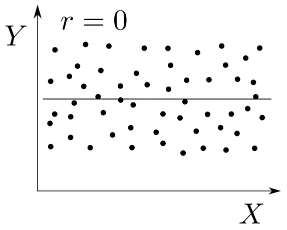


FIGURE 3.6: *No correlation.*

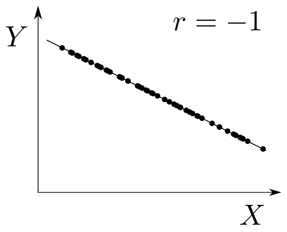


FIGURE 3.7: *Perfect negative correlation.*

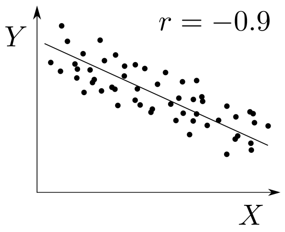


FIGURE 3.8: *High negative correlation.*

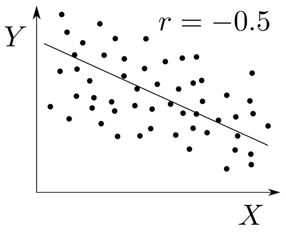


FIGURE 3.9: *Low negative correlation.*

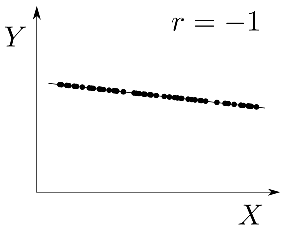


FIGURE 3.10: *Perfect negative correlation.*

Hypothesis testing

Hypothesis testing is a fundamental concept of statical analysis. This kind of test entails putting forward a hypothesis (making an assumption) about a parameter of one or more populations, which may or may not be true. Ideally the data of the entire applicable population should be included in the analysis, but since this is generally impractical, a random sample is typically used in hypothesis testing, which is assumed to represent the population [154, p. 291]. According to Weisstein [243] the process of hypothesis testing involves four main steps. These steps are described in the remainder of this section.

The procedure of hypotheses testing is comparable with that of a law trial, in which the defendant is considered *innocent until proven guilty*. The trail therefore has two possible outcomes. Two hypotheses are similarly put forward as the first step of hypothesis testing, of which only one may be considered true. These two hypotheses are known as the *null hypothesis*, denoted by H_0, and the *alternative hypothesis*, denoted by H_a, both of which need to be formulated clearly [99, p. 67].

Consider the law case analogy from the perspective of the prosecutors, for whom the desired outcome is that the defendant is found guilty. According Montgomery and Runger [154, p. 291], H_0 is stated in such a manner that it contradicts the theory the researcher would like prove, and thus H_a is stated such that it is in agreement with the theory the researcher is attempting to prove. In the analogy, H_0 represents the verdict or outcome of *not guilty*, while H_a represents the actually desired outcome of *guilty*. At the start, H_0 is assumed to be true, while H_a is eventually only accepted if sufficient evidence against H_0 is found. The reason for this approach is that one would like to avoid convicting an actually innocent defendant, and thus the approach leans on the side of caution. If a truly innocent defendant is convicted, this mistake is known as a type I error, while if a truly guilty defendant is found innocent, the mistake is known as a type II error, as illustrated in Table 3.1. Note that the actually correct outcome cannot be known unless the entire population has been examined, which is usually an impractical task, as mentioned.

TABLE 3.1: *Possible outcomes of hypothesis tests, adapted from Montgomery and Runger [154, p. 293].*

		Actual outcome:	
		H_0 is true	H_0 is false
Predicted outcome:	Fail to reject H_0	(a) True positive	(b) Type II error
	Reject H_0	(c) Type I error	(d) True negative

It is vital to note that H_0 may only be either rejected or not rejected — it cannot be accepted. That is, the defendant may be found innocent in the analogy, but not without doubt. Alternatively put, the defendant may only be found guilty based on a lack of evidence against H_0, not based on sufficient evidence for H_0 [154, p. 292].

The second step in hypothesis testing identified by Weisstein [243] is that an appropriate test statistic has to be identified and utilised so as to assess the evidence associated with the null hypothesis. The test statistic is obtained by applying a statistical test to the sample data. Examples of such statistical tests include the *likelihood-ratio test*, the *Lagrange multiplier test* or *score test*, and the *Wald statistic test* (to be discussed in more detail later).

The third step outlined by Weisstein [243] is to determine the so-called *p-values* associated with each test statistic [154, p. 292].

Thereafter, in the fourth step, once the p-values have been determined, they may be interpreted. According to Poole [180, p. 291], the p-value is the probability, assuming that the parameter specified in H_0 is indeed the true value (*i.e.* H_0 is true), of observing a result at least as far from a specified value as was observed or obtained. For example, a p-value of 3% indicates that the probability of obtaining the estimate for the parameter which was obtained is 3%, assuming H_0 is true. In the law trail scenario above, Greenhalgh [81, p. 422] and Wackerly *et al.* [240, p. 513] state that the p-value may be seen as the amount of evidence for the H_0 case, in the sense that as the p-value decreases, the evidence for H_0 decreases and the evidence for the H_a increases. Once this latter evidence increases beyond a certain threshold, *i.e.* the p-value decreases below a certain cut-off point, H_0 is rejected (the defendant is found guilty in the analogy). This cut-off point is known as the *significance level* of the test and is denoted by α. If, however, the evidence remains below a cut-off point specified, *i.e.* the p-value remains above α, H_0 is not rejected [83, p. 143], [116, p. 386], [154, p. 293, p. 300].

The traditionally acceptable value for α is 5%, but this value may be taken as low as 1% (or even less) for cases where more certainty and confidence is required, such as in medical studies. By selecting an α-value of 5%, what is in fact being pursued by the researcher is a statement of the form "We are willing to take the risk that there is a 5% probability of rejecting H_0, even if it is actually true" (*i.e.* of making a type I error) [99, p. 67], [154, p. 293], [246, p. 1320].

Based in the value of α, it is possible to determine a *confidence interval* (CI) as CI $= (100-\alpha)\%$, which is a measure of the reliability of an estimate of the parameter of interest [83, p. 143], [99,

p. 66]. According to Poole [180, p. 292], by drawing many random samples from the population and setting up a 95% CI (it may also be a different value), it may be said that '95% of those samples will contain the true parameter of interest.' Keep in mind that the parameter of interest for the population is not a random variable, but because of the sampling, randomness is introduced to the parameters of individual CIs. Thus, *before* a CI has been drawn up, it may be said that 'if I were to determine a 95% CI for a random sample from the population, there would be a 95% probability that the true parameter of interest is contained within that CI.' It is possible to state the above since the true parameter of interest, although unknown, is a constant, but the CI is still random since it has not yet been determined. Notice that at this point the above statement refers to a random sample, it does not refer to a specific sample or its CI.

Poole [180, p. 292] points out that it is vital to note that once the CI has been determined, however, the above statement will be incorrect and is a common misconception in the literature. This is because once the CI has been determined, both the CI and true parameter of interest are constants and not random. Thus, when interpreting a specific and known CI it may be stated that 'there exists 95% confidence that this CI contains the true parameter of interest.'

Hence there exists two measures for determining whether or not H_0 should be rejected: Investigating whether the p-value is less than α, or investigating whether the parameter of interest (*e.g.* $\hat{\beta}_1$) lies outside the CI. These two measures of statistical significance will always agree [51, p. 336], [83, p. 143], [116, p. 390], [186, p. 337], [240, p. 511].

Although these two methods provide the same information in respect of statistical significance, the CI is the more informative due to the following three trains of reasoning by de Prel *et al.* [51, p. 336] and Gupta [83, p. 144]. First, the CI is the interval in which the true value of the parameter of interest lies with $(100 - \alpha)\%$ confidence and secondly it is provided in the same measurement unit as was stated in H_0. An analyst will therefore have a better idea of the range (in understandable units) of the true value of the parameter with $(100 - \alpha)\%$ confidence. This is vital as an organisation might, for example, only wish to introduce a specific factor if they can be reasonably confident that it will result in a certain minimum improvement. The lower bound of the CI will make it possible to present the expected minimum improvement by the introduction of the factor. Thirdly, the CI provides information as to the direction of the effect. If, for example, H_0 is concerned with the slope parameter of a regression line β_1 and the CI lies entirely on one side of zero, the slope direction is known with a certain probability. The fourth advantage, noted by Gupta [83, p. 144] and Poole [180, p. 291], is that a CI allows for the precision of the estimates to be gauged by the width of the CI, as a narrower CI indicates more precision than a wider CI. This is information that the p-value cannot provide, and thus it is important to note that a small p-value does not indicate a narrow CI and *visa versa*. Variables identified as having more precision are those that are least affected by chance (least influenced by random error). It would thus be foolish only to consider p-values without their CI counterparts. A warning is issued by de Prel *et al.* [51, p. 336] though, that the larger a sample, the narrower the CI will be and *visa versa* for small sample [116, p. 390].

According to Montgomery and Runger [154, p. 293–294], a general misconception in the literature is that the p-value is the probability of making a type I error. It is not — the α-value is that probability. The p-value is rather the probability that a result at least as extreme would have occurred randomly. In other words, assuming that H_0 is true (as indicated above), the probability of randomly obtaining the parameter results which were obtained from the sample (or more extreme results) is the p-value. Simply put, the p-value is the probability that the observed outcome has occurred by chance. It should become evident that this ties back in with a how a decreasing p-value is equivalent to increasing evidence against H_0.

Predictive accuracy

Suppose there is interest in discovering whether or not the independent variable X in (3.2) is significant in predicting the outcome of the dependent variable Y (*i.e.* whether a change in the value of X results in a change in Y). This may be achieved by considering the coefficient β_1, which is associated with X, since if the value of β_1 does not equal zero, it would imply that there is a correlation present between X and Y, however small it might be. In order to assess this question the null and alternative hypotheses may be taken as

$$H_0 : \beta_1 = 0$$

and

$$H_a : \beta_1 \neq 0,$$

respectively.

By rejecting H_0, it is confirmed that β_1 does not have a slope of zero, and thus the independent variable associated with β_1 is statistically significant in the sense that, to some degree, there exists correlation between the variable associated with β_1, X that is, and the output variable Y. Thus, using X it should be possible to produce predictions for Y, but note that, as stated above, and illustrated in Figures 3.3–3.5, the sign of β_1 does not indicate the level of correlation, and thus predictive ability, of the independent variable X [116, p. 376], [128].

Some authors, like Chatterjee and Hadi [40, p. 39], state that a small p-value automatically indicates a strong correlational relationship, but other authors such as Lane *et al.* [116, p. 376], caution against such rushed judgements. They believe that a small p-value is reason to suspect correlation, but it does not reveal the magnitude of the correlation and that any p-value results should be considered in conjunction with the CI, the β-value of and the predictive accuracy of the model containing the independent variable of the β-value in question. It is therefore necessary to test the predictive ability of a model containing that variable in the context of new and previously unseen data. A detailed discussion on assessing the predictive ability of a model follows later in this literature review.

Analyses of large samples are known to produce small p-values and narrow CIs more easily, and small samples the reverse. These two phenomena should be noted while interpreting p-value and CI results. For example, a significant p-value and a narrow CI from a small sample is much more significant, and *vice versa* [51, p. 336], [83, p. 143], [220, p. 279].

Multiple linear regression

Up to now the discussion has focussed on models with only a single independent variable. According to Winston [246, p. 1318], *multiple linear regression* is an extension of linear regression arrived at by the addition of one or more independent variables, but still allowing only one dependent variable, such that (3.2) becomes

$$E(Y_i \mid X_{1,i}, X_{2,i}, \cdots, X_{m,i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}, + \cdots + \beta_m X_{m,i}, \quad i \in \mathcal{N}, \tag{3.7}$$

where the m independent variable values in the i-th observations of the sample are denoted by $X_{1,i}, X_{2,i}, \ldots, X_{m,i}$ and their corresponding coefficients by $\beta_1, \beta_2, \ldots, \beta_m$, respectively.

3.1.4 The notions of explanation *versus* prediction

Before continuing it is important to distinguish between two different types of modelling, known as *explanatory modelling* (a subset of inference) and *predictive modelling* [204, p. 291]. The aim in explanatory modelling is to test causal hypotheses by application of statistical models. Confirmation is therefore desired that the factors measured by the independent variables may be assumed to cause an underlying effect in the factors measured by the dependent variables. Predictive modelling, on the other hand, involves the application of data mining algorithms or statistical models to data so as to predict the outcome of new observations. Shmueli [204, p. 291] claims that the "wrong" model in an explanatory sense may often predict better than a correct one.

Shmueli [204, p. 289] believes it a major concern that the terms predictive modelling and explanatory modelling are often combined in the statistical literature, since a lack of distinguishing between the two terms may result in the erroneous assumption that predictive power implies explanatory power and *visa versa*. This highlights the importance of defining the aim of a study, and understanding the nature and requirements of the client for whom the results are indented. Academic research is largely concerned with explanation, while industries (such as the finance industry) are generally more focussed on obtaining good predictions.

3.1.5 Methods of estimating the function $\boldsymbol{f}$

An example of the process of estimating a specific form of the function $\boldsymbol{f}$ in (3.1) has been given above. In this section, the process is further discussed in general and further examples given.

The majority of statistical learning methods used to estimate the function $\boldsymbol{f}$ in (3.1) fall into one of two categories, namely *parametric methods* and *non-parametric methods*. When deciding which method to use in the estimation, the first step is to establish whether the goal in the estimation is prediction, inference or both [99, p. 21].

When using parametric methods, assumptions are made about the shape of $\boldsymbol{f}$. No such explicit assumptions are, however, made when using non-parametric methods [99, p. 23]. The validity of this statement is, to some extent, debatable since authors such as Marascuilo and McSweeney [135] have noted that it is wrong to believe that no assumptions are made in non-parametric tests, but rather that considerably fewer assumptions and often no assumptions are made.

According to Sheskin [201, p. 97] the choice of selecting an estimation approach should be based on the type of data being evaluated. If the data are in nominal or ordinal form, the use of non-parametric methods is recommended, else if the data are in continuous form he recommends using parametric methods [154, p. 619]. While many authors believe that continuous data are more valuable than nominal or ordinal data and prefer always to use continuous data, if available, in conjunction with the best parametric test, other authors (the majority according to Sheskin [201, p. 97]) believe that if any assumptions about continuous data are violated it is best to convert the data to nominal or ordinal form and to adopt a non-parametric approach. Sheskin [201, p. 97] is only in favour of using a parametric test if no assumptions are violated about continuous data.

Parametric methods of estimation

The parametric approach, also known as the *model-based approach*, allows for the problem to be simplified from one that requires a function $\boldsymbol{f}$ in (3.1) to be estimated, to one that only requires estimation of a set of parameters. It is generally accepted that it is easier to estimate a set of parameters than estimating an entire function $\boldsymbol{f}$. The linear regression and multiple linear regression models (described in §3.1.3) are examples of parametric models since the form of the function that describes the relationship is known [99, p. 22].

James *et al.* [99, p. 22], however, claim that a major disadvantage of the parametric approach is that it often leads to a model being chosen for $\boldsymbol{f}$ that does not provide a good representation of $\boldsymbol{f}$, hence yielding a poor estimation. A possible solution to this problem is to select a more *flexible* model which allows one of a large variety of functional forms to be fitted to $\boldsymbol{f}$. This may, however, lead to a problem known as *over-fitting*, where the model chosen follows the data points, and subsequently the errors in the data, so closely that the chosen model is only applicable to the specific data set and cannot be used to estimate outputs from other related, similar data sets [127, p. 7].

James *et al.* [99, p. 22] provide the following two-step approach toward implementing the parametric method. The first step is to make an assumption about the shape of $\boldsymbol{f}$, for example assuming the multiple linear regression form of (3.7) such that

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}, \ldots, \beta_m X_{m,i}, \quad i \in \mathcal{N}. \tag{3.8}$$

Therefore, instead of having to estimate the entire function $f(\boldsymbol{X})$, the only parameters to be estimated are $\beta_0, \beta_1, \ldots, \beta_m$.

The second step is to fit the model to the data set. This step requires that values for the parameters $\beta_0, \beta_1, \ldots, \beta_m$ have to be found such that

$$Y \approx \hat{f}(\boldsymbol{X}) \approx \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}, \ldots, \beta_m X_{m,i}, \quad i \in \mathcal{N}. \tag{3.9}$$

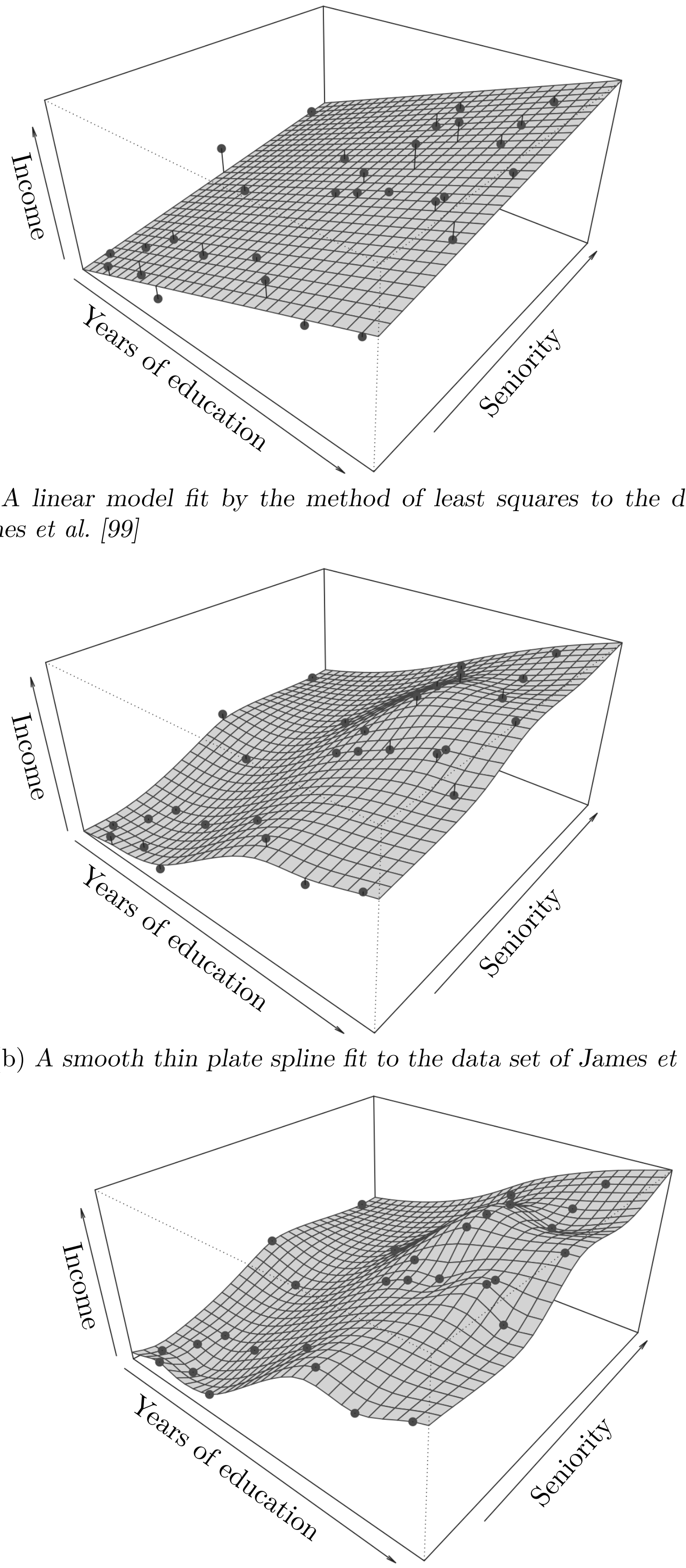
for every observation of the form $(X_{1,i}, \ldots, X_{m,i}, Y_i)$ [88, p. 28], [99, p. 21].

Data on thirty individuals were collected by James *et al.* [99, p. 22] in terms of their monthly income, the number of years they were educated, and their seniority (their age). Displayed in each of the Figures 3.11(a)–3.11(c) is a plot of the individuals' income as a function of the number of years they were educated and their seniority. These points are displayed as dark grey dots in Figure 3.11. Each of the light grey surfaces in the figures represents a specific fit with respect to the data. In Figure 3.11(a), the light grey plane represents a least squares fit of the data, while the light grey surface in Figure 3.11(b) (Figure 3.11(c), respectively) represents a smooth (rough, respectively) thin plate spline fit of the data.

The least squares fit in Figure 3.11(a) captures some of the important characteristics of the data, such as the positive relationship between the years of education and income, and also the positive relationship between the seniority and income. The fit is, however, not ideal and may lead to some inadequate estimations [99, p. 22].

Non-parametric methods of estimation

As stated above, the parametric approach allows for the estimation problem to be simplified from one that requires the function $\boldsymbol{f}$ in (3.1) to be estimated, to one that only requires estimation of a set of parameters of $\boldsymbol{f}$ — typically under the assumption of a certain form for $\boldsymbol{f}$ [154, p. 619]. This should, however, be done carefully as following the data points too closely may result in an estimation that is only applicable to the data set on hand. The advantage of non-parametric methods is that they do not necessarily become thus trapped pursuing a specific functional form, as is often the case with parametric methods. Instead, they are free to explore a much wider range of possible functional forms for $\boldsymbol{f}$. This freedom does, however, have its disadvantages, the main one being that it results in a much larger number of parameters that need to be estimated



(a) *A linear model fit by the method of least squares to the data set of James et al. [99]*

(b) *A smooth thin plate spline fit to the data set of James et al. [99]*

(c) *A rough thin plate spline fit to the data set of James et al. [99]*

FIGURE 3.11: *Three estimates of income as a function of years of education and seniority, adapted from James et al. [99, p. 23].*

as opposed to only a few pre-specified parameters when a specific functional form is pursued according to a parametric method [99, p. 23].

To illustrate the working of one of the well-known non-parametric methods, the *thin plate spline* technique, refer again to the data set of James *et al.* [99, p. 23] mentioned above. The result of estimating the functional form of $\boldsymbol{f}$ using the thin plate spline technique is illustrated in Figure 3.11(b), where the light grey surface is the estimation of $\boldsymbol{f}$. Note how well the estimation fits the true data without following it too closely. This is because the level of smoothness associated with the thin plate spline technique was selected as smooth. Montgomery and Runger [154, p. 619] claim that most non-parametric approaches have this setting available, known as the *level of smoothness*, the *level of significance* $100(\alpha)\%$, or the *confidence level* $100(1-\alpha)\%$.

If, however, this smoothness level is altered to a lower setting of smoothness, the result is the over-fitted estimation shown in Figure 3.11(c). Thus, again, the need is clear the for the data analyst to exercise discretion with respect to choosing the level of smoothness.

3.1.6 Trade-offs between prediction accuracy and model interpretability

When choosing which approach to use for estimating the function $\boldsymbol{f}$ in (3.1), the trade-off between the accuracy of the estimation and how interpretable the model is should be considered. The one extreme is a very simple and rigid approach, such as linear regression, which does not allow for much variation with respect to the functional forms that can be used to estimate $\boldsymbol{f}$. Although parametric estimation methods, such as linear regression, allow for limited functional shape options, the simplicity of the methods typically admit easy interpretation, and are therefore often favoured in inference analyses. The other extreme predominately occurs in prediction where the preference shifts to flexibility at the cost of more challenging interpretability [99, p. 25]. The two extremes mentioned above are well represented in Figure 3.11, where Figure 3.11(a) contains the results of a less flexible, but intuitively interpretable estimation, whereas the result in Figure 3.11(c) is a more flexible estimation which is harder to interpret.

3.1.7 Supervised *versus* unsupervised statistical learning

Hastie *et al.* [88, p. 29] describe how *supervised* learning models attempt to *teach* the function $\boldsymbol{f}$ by making use of the response variables in equations, while James *et al.* [99, p. 26] explain that in *unsupervised* learning, the analyst effectively works blind, because he only has access to the measurement variables that go into an equation, but not the response variables. The majority of statistical learning problems can be classified as one of these two methods.

The main goal in supervised learning, the paradigm of focus in this thesis, is that once the learning process has been completed, the outputs from the then 'educated' system should be of such a standard that it should adequately represent actual outputs for all other data sets that are likely to be used [88, p. 29]. All of the methods discussed above fall into this category.

It is natural to question the use of unsupervised learning methods, a field much less developed in the literature. The answer given by James *et al.* [99, p. 27 & 373] is that this learning paradigm is very useful if one wishes to better understand the relationship between the variables and/or observations. An example of such a method is *clustering*, which aims to sort a data set into distinct categories.

3.1.8 Regression *versus* classification problems

The two main categories of data, qualitative data and quantitative data, are very important in the context of classifying the outputs of a process. Hastie *et al.* [88, p. 10] claim that attempts at predicting outputs of a quantitative nature are collectively referred to as *regression*, while attempts at predicting qualitative outputs it is referred to as *classification*. Various other authors concur with this taxonomy [75, p. 79], [127, p. 8], [199, p. 223]. James *et al.* [99, p. 29] echo this differentiation by noting that the response variables, not the predictor variables, should be the main consideration when selecting between a regression or classification approach to tackle a particular statistical learning problem.

3.1.9 Types of research studies

According to Lund and Lund [128], there exist two main experimental study types, namely *experimental research* (or *intervention research*) and *non-experimental research* (or *observational research*). Experimental research refers to research during which the independent variable(s) of the study are manipulated as to analyse the effect that this has on the dependent variable(s), such as in field trials. Non-experimental research on the other hand, according to Silva [206, p. 83], does not entail the manipulation of the independent variable(s), since it is either impossible, impracticable or unethical. Rather, past data are observed, which may include data realising without the direct influence of the analyst or researcher, even if freshly collected using surveys. Cohort and survey studies reside in the class of observational research.

3.1.10 Internal validation *versus* external validation

Recall, from §2.4.3, the concepts of *verification* and *validation*. Verification is the process of ensuring that a model performs as expected, *e.g.* are the data imported as expected, is the correct test performed on the data, and are the results as expected. Validation, on the other hand is the process of assessing whether or not a model adequately and accurately represents the real world, *e.g.* may it be deduced with confidence that the results obtained from the model reflect what actually happens in reality [117, p. 333].

Luna [127, p. 6] is of the opinion that the predictive ability of statistical models is generally overestimated. This occurs due to the fact that normally a sample data set is used to build and then later assess the model. In the context of statistical tests, one may validate the model by considering the predictability of the model by using previously unused data. In this manner, one may be able to ascertain whether or not the model created using only a sample of data may be seen as adequately representing the population. The preferred option, as mentioned above, is thus to use a second set of data, which may be done in one of two main ways [170, p. 160].

The first, *internal validation*, refers to only having access to one sample of the population. In which case a sub-sample of that sample may be used to build the model which is then tested against the remaining data of the sample. Internal validation is popular when only smaller samples are available. Internal validation may be achieved by partitioning the only available data set as shown in Figure 3.12, where a specific percentage of the observations are set aside for eventual validation, and the remaining observations are used for learning.

The second option, *external validation*, occurs when a model is developed using a sample of the population and tested on another, independent, sample from the same or similar population. The context of the second sample should, however, be considered as it may lead to poor results and thus produce a lack of confidence in the model when, in fact, it should not be the case. With reference to Figure 3.12, external validation may be achieved by using one entire data set for learning, and a second, external data set for the validation [170, p. 161].

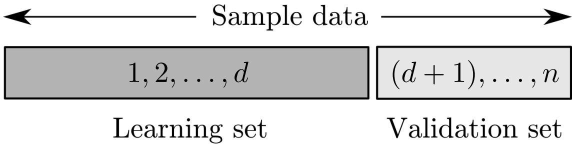


FIGURE 3.12: *Partition of learning and validation subsets.*

3.1.11 Data splitting methods

For the purposes of enhancing the learning of models, as discussed above in §3.1.10, methods of *data splitting* or *re-sampling* may be used [99, p. 175]. Two such methods, discussed below, are *cross-validation* (CV) and *bootstrapping*.

Cross-validation

Two main methods of CV exist, *K-fold cross-validation* and *"leave-one-out" cross-validation*. K-fold CV involves using only a fraction $100\left(\frac{K-1}{K}\right)\%$ of the data set to learn the models, while the remaining fraction is used to test for misclassification. This step is repeated K times, for a different value of K, and thus a different, but mainly the same $100\left(\frac{K-1}{K}\right)\%$. The next step is then to use the entire data set to build the model. The misclassification or error rate of this final model is then estimated as the average of the K misclassification rates determined during the previous step. The process of K-fold CV is depicted in Figure 3.13. The most popular K-fold CV method is the well known *10-fold* CV [127, p. 7].

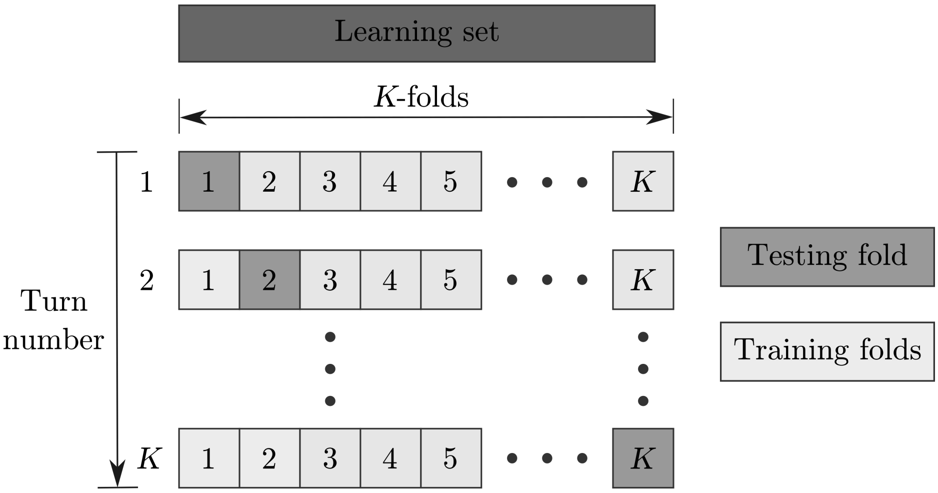


FIGURE 3.13: *The notion of K-fold cross-validation.*

In the second method of "leave-one-out" CV a very similar approach is adopted. All the observations in the data set, except one, are used to build the model, and the model is then tested by considering the one remaining datum. This process is repeated as many times as there are observations in the data set. The final model is constructed using the entire data set and the error rate is estimated from all the error rates of the individual data points against the various models [127, p. 7].

Bootstrapping

A second type of data splitting method is known as *bootstrapping* or *bootstrap re-sampling*. As in CV, the idea is to create subsets of the learning data of a specific size, but the method in which these subsets are obtained differ. Bootstrapping, according to Tonidandel *et al.* [227, p. 390], involves repeatedly and with replacement selecting observations from the learning data set at random to form the subsets. Unlike CV, in which the number of subsets of the data set is limited to the number of folds and the size of the learning data set, since no observation may belong to more than one testing fold, this restriction does not apply to bootstrapping — the number of testing folds may, in fact, be as large as the user would like. The process of bootstrapping is illustrated in Figure 3.14, for B bootstrap re-sampling rounds. After each round, the data are 'returned' to the learning dataset and the process is repeated. This process is known as *random sampling with replacement*.

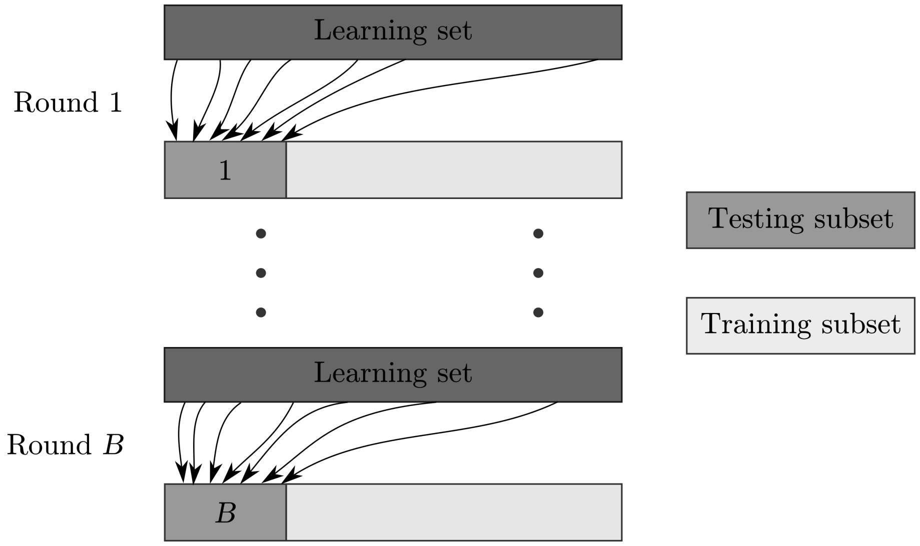


Figure 3.14: *The notion of bootstrapping.*

3.2 Reasons for correlational relationships

Trochim [229] describes a correlated relationship as two variables in harmony, meaning that as the one variable changes, the second variable adjusts in a proportional manner, based on some function. The notion of correlation was touched upon in §3.1.3.

Once a correlated relationship has been identified between two variables, it is necessary to investigate why this relationship exists. The natural assumption is that correlation implies causation, but rushing to such a conclusion is considered dangerous by Silva [206, p. 277], who states that a correlated relationship may be due to one of the following four reasons: measurement error, chance, confounding variable(s), or a causal relationship. These notions are elaborated upon in this section.

3.2.1 Bias caused by measurement or selection error

The concepts of measurement error and selection error are described in this section, as well as methods of avoiding or reducing the effect of such errors.

Measurement error

According to Trochim [228], measurement errors in data may cause the means of the samples to shift up or down due to the *bias* introduced by these errors. As the name suggests, *measurement errors* relate to the errors that have infiltrated into the data during the process of collecting, observing and/or measuring the data. Silva [206, p. 282] states that this type of error may occur within data by one of three ways, namely by the observer, by the subject or event of interest, or by the instruments used to collect the data. The observer is the person who observes the subject or event and records the necessary findings, whereas the subject or event of interest is what is being observed and studied. The instruments refer to the methods used to make the measurement, such as a questionnaire or a wind speed reader. Silva [206, p. 285] and Trochim [228] provide the following advice in order to reduce measurement errors:

1. Any data being typed into a computer or written onto a form by an observer should either be double-checked by a second party or entered twice by the same individual.
2. Anyone involved in the data collection process should be adequately trained.
3. No observer who might be biased should be allowed to participate.
4. No subject who might be biased should be allowed to participate.
5. The accuracy of any physical measuring instruments should be verified.
6. If surveys or questionnaires are used, they should be designed in such a way so as not to bias participants.
7. Clear protocols should exist for the collection of data so that the data collection process can be carried out in a consistent manner.
8. Any information provided by subjects should be subjected to verification.

Selection error

Selection error is another form of bias which ties in with measurement error — specifically with how measurement errors are introduced into the data through the subjects or events of interest. Selection error, according to Silva [206, p. 278], refers to a biased data set that occurs when only a specific subset of participants or subjects from the population reported to be represented are selected to collect data on, or from. This implies that the final data set contains only a certain group of subjects, or excludes a certain group of subjects, from the much larger group who are eligible to be selected for the study. The cause of errors in this case may be that the selected group of subjects have some characteristic in common which resulted in them being selected while other subjects, who were also eligible for selection, are excluded as a result of not exhibiting this characteristic.

Consider the example of a survey which is made available to young people of a certain age group and is concerned with the health of their lifestyles. Assume that it is also communicated beforehand that they would receive feedback from experts. It is possible that the majority of people who then volunteer to participate in the survey have a reasonably healthy lifestyle of which they are not ashamed, whereas people who do not have a healthy lifestyle might not have been as excited about the prospect of participating in the survey as they might feel that they are subjecting themselves to scrutiny by the experts. Incorrect conclusions may then be drawn about the lifestyle of the average young person of that age category.

In order to avoid selection error, Silva [206, p. 281] suggests that it is essential that the method used to sample the participants should be chosen judiciously. The sample selected should accurately represent the population from which it is selected. If historical cohort data are used, for example, thought should be given to what biases may be present in the existing data and how the data can be altered so as to exclude it.

3.2.2 Chance relationships

The possibility also exists that a collinear relationship between two variables is simply due to chance. To gain more surety that the collinearity is not only due to chance, the statistical significance of the relationship may be investigated by considering the corresponding p-value. If the p-value is well below the typical cut-off value of 0.05, then the possibility of the relationship being purely chance is small. As stated above in §3.1.3, however, small relationships which are in reality insignificant may show up as significant in large samples. The opposite is true for small samples, *i.e.* where relationships that are in reality significant show up as insignificant [206, p. 294].

3.2.3 Confounders

Another possible reason for collinearity is the existence of a *confounding variable* or a *third variable*. A confounding variable is a variable that is currently not in the model, but is in a correlational and causal relationship with both the independent and dependent variables currently in the model [206, p. 291], [228].

Consider, for example, a regression model concerned with the number of ice creams sold on a specific beach each day and the number of drownings that occur at that beach. Suppose these two variables are highly correlated. A confounding variable in this example may be temperature, because on warmer days more people are at the beach and hence more ice creams are sold, and the probability of someone drowning is high. This type of situation regularly occurs when regression is misused to provide false results which state that two variables are in a causal relationship when, in fact, they are only correlated. For this reason, the phrase *correlation does not imply causation* has become popular.

Silva [206, p. 313] suggests that a good starting point to search for confounders is to consider the findings of similar studies previously completed. Examples of typical confounders in studies about subjects include their socio-economic status, geographical information, race, gender, age and in most health-related studies, smoking status. Harrell [86] nevertheless notes that the final decision as to whether or not a specific regression model includes possible confounders lies largely with the analyst. This, again, emphasises the need for human expertise and sound personal judgement throughout the study and that the results from any automated procedures should be considered with care.

According to Silva [206, p. 293], confounders may be dealt with either at the design stage of the study or later during the analysis phase, but only once the relevant data have been collected. This emphasises the need for a study to be well designed so that data may have been collected on those potential confounders from the start. The three approaches that may be adopted to control confounders are *randomisation*, *restriction*, and *matching*.

The approach of *randomisation* entails randomly selecting and partitioning data points or subjects into classes. This will most likely result in known and unknown confounders being distributed evenly among the different sample classes in the study being compared as well as among those subjects randomly chosen not to be part of the study although they did, in fact, qualify for inclusion in the study. Randomisation is only applicable in experimental studies, as discussed in §3.1.9.

The approach of *restriction*, involves ensuring that only subjects which have a similar relation to any suspected confounders are included in the study. For example, in a study in which the success of students at tertiary level is investigated by considering personal and high school academic factors, a possible confounder may be the type of home in which they grew up. One possible way of ensuring that this potential confounder does not distort the data is to include only students who come from homes associated with an average income which falls within a certain bracket.

Matching is only applicable to studies where subject or data points exist within the control and exposure groups, such as in medical studies. It involves matching or pairing subjects from the control group with subjects in the exposure group who exhibit the same value for a certain variable selected to be the matching variable. It is possible that such matching variables may be confounders. Marsh *et al.* [137, p. 327] state that if the suspected confounder is, in fact, only correlated with the independent variable, the unwanted effect of *overmatching* may take place. Another criticism of matching is that since one may not use the variable selected to match with the data in the model, information is lost.

If data collection has already been completed, the technique of *stratification* may be used to control confounding during the analysis. *Stratification* can best be explained by means of an example. Suppose that a certain study examines whether people who eat more than a certain amount of peanuts a day are fitter than those who do not. Whereas it may well be found that those who eat more peanuts are fitter, it is reasonable to suspect that the number of hours exercised per week may be a confounder. Exercise may be considered a confounder if it predicts both how many peanuts are eaten by subjects and their fitness level. The process of *stratification* would then involve separation of the subjects based on the number of hours a week they exercise, after which the analysis should be repeated for each category. The suggestion by Silva [206, p. 294] is then to weight the results from the different tests of each category and combine them.

3.2.4 Causal relationships

If it would seem that a correlational relationship is not due to bias because of selection or measurement error, chance, or confounders, it may be assumed that the relationship is causal. According to Trochim [228], a *causal relationship* is one in which the change in one variable directly causes another variable to alter as well. Returning to the example of ice cream sales at a beach, inclusion of temperature as a variable in the model may cause an increase or decrease in ice cream sales.

Although it is highly unlikely to prove conclusively that a correlation is due to bias in the selection or measurement error, chance, or confounders, Silva [206, p. 286] has provided a list of eight aspects that may be considered so as to strengthen the argument that the relationship is indeed causal. These aspects should not be interpreted as criteria for causality to exist, or whose absence proves that it does not. Instead, the presence of these criteria should be interpreted as additional evidence that a causal relationship may exist. The only exception to this is the first aspect, namely a *temporal relationship*, which has to be present for a causal relationship to exist. The eight aspects are a *temporal relationship*, *biological plausibility*, *consistency*, *strength*, an *exposure-response relationship*, *specificity*, *reversibility*, and *coherence*. These aspects are described briefly in the remainder of this section.

A *temporal relationship*, requires that an event linked to the independent variable should precede the event of the dependent variable. This can easily be confirmed or denied using cohort data.

In the medical field, *biological plausibility* is present when the finding of an analysis is consistent with existing biological knowledge. The lack of such a finding should not be seen as too detrimental to the cause of proving causality, as it may well be the case that knowledge which suggests that the relationship is plausible is lacking in the literature.

Consistency occurs if similar results to the current analysis have been found by similar studies, utilising different populations. A lack of consistency does not necessarily point to non-causal relationships.

If the *strength* of the relationships between variables are found to be particularly strong, it is much less likely that those relationships are a result of chance, bias, or confounders.

The *exposure-response relationship* aspect is only applicable to specific scenarios in which the occurrence or event of an independent variable, as well as the exposure time of that variable, may be measured. If it can be shown that an increase in the exposure results in a direct change in the dependent variable, this finding may be used as additional evidence of possible causality.

The aspect of *specificity* is relevant when different dependent variables are being considered and it is found that a change in a specific independent variable results in a clear change in one of the dependent variables, but none of the others. Specificity in this context should not be confused with the notion of true negative rate of predictions as discussed in a later section.

Reversibility may be tested by removing an independent variable which is a potential cause of the dependent variable and ascertaining whether the response by the dependent variable is reduced.

The last of the eight aspects, *coherence*, requires that the findings of a study should not strongly conflict with proven facts in the literature related to the topic of the study.

3.3 Why is linear regression not appropriate for classification?

For illustration purposes example data are used in this section in conjunction with an example study to showcase how classification typically works and why linear regression is not appropriate for classification problems. The data set is shown graphically in Figure 3.15. The data relate to individuals who have defaulted on their credit card payments, and are displayed in scatter diagram format with the individuals' monthly incomes as a function of their credit card balances. Those who have defaulted on their credit card payments are displayed as crosses and those who have not are shown as circles.

It should be evident, upon scrutiny of Figure 3.15, that the individuals' income values are not good indicators of who will default, while their credit balances do a much better job in this respect. This is an example of a rather simplistic nature, but an ability to notice the different groups at a glance is not normal in most cases.

With reference to the traditional multiple linear regression form in (3.7), the response variable, Y, would be individuals who default, and the independent variables will be their credit balances, X_1, and income, X_2. The method of linear regression discussed in §3.1.3, however, is now not appropriate since Y is not quantitative.

To use linear regression in conjunction with a qualitative output would require the assignment of quantitative values for those potential outputs, and this is where the problem arises. For example, if the objective is to predict what sport people play based on their muscle mass in specific body regions, the possible output Y could be

$$Y = \begin{cases} 1 & \text{if cricket,} \\ 2 & \text{if rowing,} \\ 3 & \text{if tennis.} \end{cases}$$

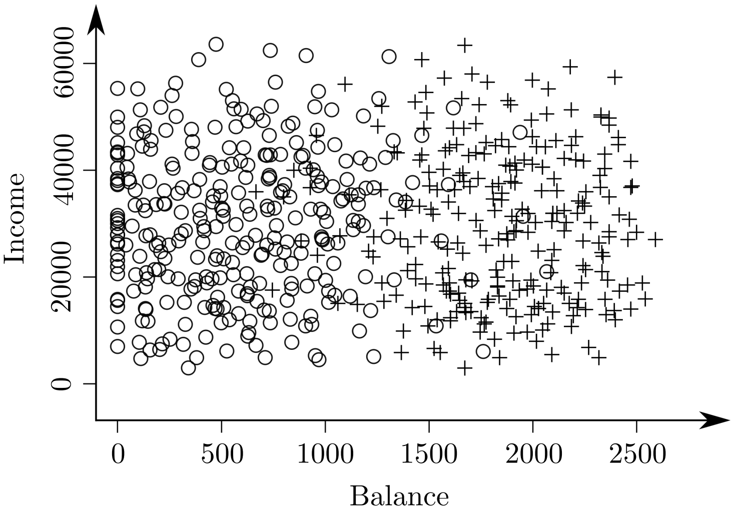


Figure 3.15: *Scatter plot of individuals who did and did not default on their credit card payments shown as the individuals' monthly incomes as a function of their credit card balances.*

There are two problems with this setup. First, it implies that the outcomes can be ordered, *i.e.* that rowing falls between cricket and tennis. Secondly, it also suggests that the difference between the outcomes are of equal size. If the order of these outcomes were to be altered, the result is that the changes suggest different relationships between the variables and the linear regression models produced will be different. The new model may lead to new predictions which underline the inconsistency of using linear regression in the context of qualitative outputs [99, p. 129]. These problems may be addressed partially if there is a natural ordering between the variables (for instance, if the outputs are low, medium and high). It is, however, rare that qualitative responses can be converted into quantitative values with an ordering and an equal distance between them.

What is potentially more practical is if the outputs are binary in nature, *i.e.* with just two possibilities. An example of such a scenario is if the objective is to predict whether or not a student has successfully graduated from their course, in which case a *dummy variable* may be employed such that the possible output could be stated as

$$Y = \begin{cases} 1 & \text{successfully graduated,} \\ 0 & \text{otherwise.} \end{cases}$$

Using linear regression it would then be possible to obtain the predicted outcome of $\hat{Y}$ which in theory should range between zero and one. By applying a rule such as "If $\hat{Y} \geq 0.5$ change to one, else zero," it is possible to convert the quantitative response to a binary value. In this manner it is easy to apply linear regression. Even if the above order of value assignment is changed, the resulting linear regression model will provide the same results.

The method of dummy variables may also be used to convert categorical independent variables with more than two categories into a form appropriate for a classification method. Consider the example of a single independent variable, called *Race* and which indicates the race of a student, being used to predict the binary outcome of whether or not a student will be successful at their tertiary studies. Assume that for a particular sample, the possible categories are available for *Race* are *Black*, *Coloured*, *Indian*, and *White*.

With the assistance of a dummy variable the above example may be altered to the following three binary variables of, [109, p. 110],

$$X_{i,1} = \begin{cases} 1 & \text{if the } i\text{-th student is Black,} \\ 0 & \text{otherwise,} \end{cases}$$
$$X_{i,2} = \begin{cases} 1 & \text{if the } i\text{-th student is Coloured,} \\ 0 & \text{otherwise,} \end{cases}$$
$$X_{i,3} = \begin{cases} 1 & \text{if the } i\text{-th student is Indian,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$X_{i,4} = \begin{cases} 1 & \text{if the } i\text{-th student is White,} \\ 0 & \text{otherwise.} \end{cases}$$

As pointed out by Starkweather [211, p. 2], however, if there exist k categories inside the original independent variable, only $k - 1$ dummy binary variables have to be created as one of the categories has to be used as the *reference class* or *baseline class*. If all the categories were to be included, the assumption of multicollinearity, to be discussed later, will be violated [109, p. 111].

Thus for the above example, let the *Black* category of *Race* be the reference class, which will lead to the model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} = \begin{cases} \beta_0 & \text{if the } i\text{th student is Black,} \\ \beta_0 + \beta_1 & \text{if the } i\text{-th student is Coloured,} \\ \beta_0 + \beta_2 & \text{if the } i\text{-th student is Indian,} \\ \beta_0 + \beta_3 & \text{if the } i\text{-th student is White.} \end{cases} \tag{3.10}$$

The interpretation of results from this example will be further expanded upon in a later section.

The selection of the reference class should, however, be done with the final interpretation of the results in mind as all the remaining categories are compared against the reference class. For studies that include a control group, its category should be selected as the reference class as it is the category against which one would like to compare the remaining categories. Some studies, however, do not have a control group. In such cases Starkweather [211, p. 2], suggest, that a category with few cases should not be used as such a category as the reference case will result in higher multicollinearity. In addition, the use of an 'extreme category' will make the interpretation of results easier. In the case of a classification model the term 'extreme category' refers to the category that has either the highest or lowest proportion of the desired outcome.

The selection of the reference class and coding method employed has no effect on the overall model fit and thus predictive ability of the model, but the individual p-values and coefficient values will change as with each different reference class the remaining classes are being compared to that new reference class. The p-value results of an overall model fit, however, such as the log likelihood to be discussed later in this thesis, is unaffected [99, p. 86].

Predicting whether an individual plays cricket using linear regression is relatively easy. Also, even if the above order of value assignment is changed, the resulting linear regression model will provide the same results. The method of least squares may furthermore also work well in this case. Despite the potential use of linear regression in the context of a binary response variable, a major problem is still that some of the estimates made using least squares could potentially fall outside the standard [0,1] interval, resulting in the interpretation of the results as probabilities being impracticable. Figure 3.16 contains an example of this phenomenon. The data set from the earlier example involving individuals and their credit card defaults was used to produce the graph.

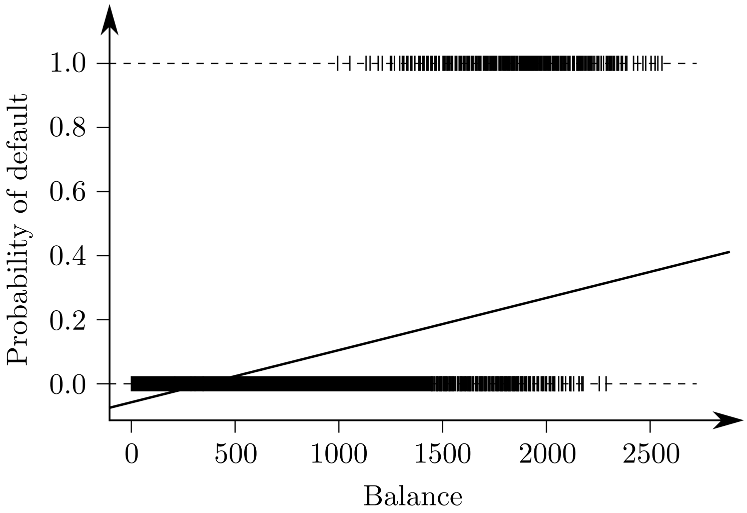


Figure 3.16: *Estimated probabilities of individuals defaulting on their credit card payments, as determined by applying linear regression to the data set in Figure 3.15.*

It is evident that specific conditions need to be present for linear regression models to be applicable for classification. This highlights how it is not ideal to use a standard estimation model for the scenario of qualitative outputs. The question then becomes: What is applicable? The answer is a classification model. Five well-known techniques which may be used for classification are described in the sections to follow.

3.4 Logistic regression

Gelman and Hill [75, p. 79] are in agreement with James *et al.* [99, p. 129] in that the use of linear regression is ill-sufficient when attempting to model binary outcomes, such as success or failure of tertiary studies. A better alternative, according to Luna [127, p. 8], is the parametric method with prediction as its main aim, known as *logistic regression.* In this paradigm, observations from a data set are classified into specific categories, and the fundamental assumption is that the probability of a specific event occurring follows a *logistic distribution.*

According to Gelman and Hill [75, p. 80], this distribution has two advantages. In the context of the example data of Figure 3.15, these advantages are illustrated in Figure 3.17.

The first advantage is the fact that the distribution in Figure 3.17 is bounded from below by zero and from above by one. This enables it to represent probabilities and avoid the problem often faced when using linear regression whereby predictions fall outside the standard [0,1] interval. Secondly, the distribution takes the form of an "S" shape, meaning that small differences in the values of the observations at the centre will have a much more pronounced effect on the outcome than similarly sized differences at the extreme values [127, p. 8].

Again consider a sample of n observations indexed by the set $\mathcal{N}$. Shaliziz [199, p. 224] explains that if an output Y is determined to be of a binary nature, logistic regression will assist in modelling the conditional probability $P(Y = 1 \mid \boldsymbol{X} = [X_1, X_2, \ldots, X_m])$ as a function of $\boldsymbol{X}$ for the case of m independent variables.

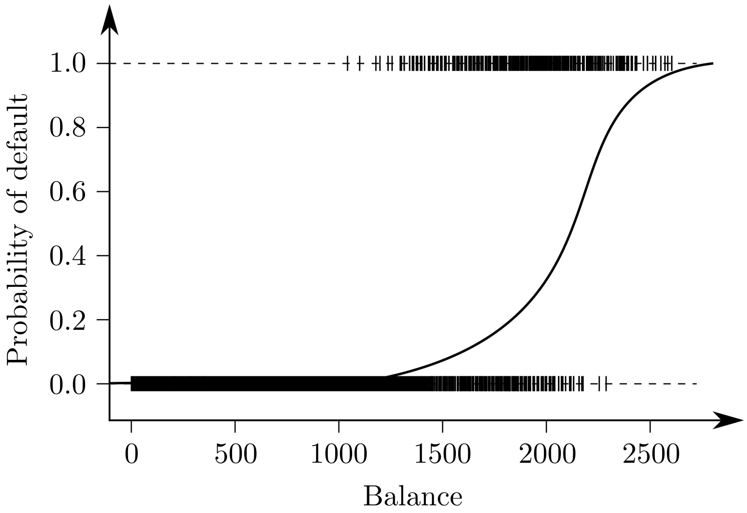


Figure 3.17: *Estimated probabilities of individuals defaulting on their credit card payments, as determined by applying logistic regression to the data set in Figure 3.15.*

3.4.1 The logistic regression model

Lekdee and Ingsrisawang [118, p. 1] express, for the case of m independent variables, the *logistic function* as

$$P(Y_i = 1 \mid \boldsymbol{X} = [X_{1,i}, X_{2,i}, \ldots, X_{m,i}]) = \frac{e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i},\ldots,\beta_m X_{m,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i},\ldots,\beta_m X_{m,i}}}, \quad i \in \mathcal{N}. \tag{3.11}$$

For ease of explanation, the single independent variable case of

$$P(Y_i \mid X_i) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}, \quad i \in \mathcal{N} \tag{3.12}$$

or, in rearranged form,

$$\frac{P(Y_i \mid X_i)}{1-P(Y_i \mid X_i)} = e^{\beta_0+\beta_1 X_i}, \quad i \in \mathcal{N} \tag{3.13}$$

is considered here. Gelman and Hill [75, p. 82] call the left hand side of (3.13) the *odds*, which may assume any value in the interval $[0, \infty)$. These odds takes the form $\frac{P}{1-P}$ for an event which has the probability P of occurring and a probability $1 - P$ of not occurring.

Consider, as an example, an event where one in five people watch sport, which implies that the odds are $\frac{1}{4}$ that someone watches sport. The reason for this is that 1 in 5 means that $p(X) = \frac{1}{5} = 0.2$, and when substituted into the left hand side of (3.13) the result is $\frac{0.2}{1-0.2} = \frac{1}{4}$ [99, p. 132].

Bowers [26, p. 103] simplifies the relationship between a probability and odds by considering the equations

$$\text{Risk} = \text{Probability} = \frac{\text{Odds}}{1+\text{Odds}}$$

and

$$\text{Odds} = \frac{\text{Probability}}{1-\text{Probability}}.$$

Therefore, odds of 0.5 and 2 are equivalent to event probabilities of $\frac{1}{3}$ and $\frac{2}{3}$, respectively. The ratio of the two odds,

$$\frac{\left(\frac{P_1}{1-P_1}\right)}{\left(\frac{P_2}{1-P_2}\right)} = \frac{\left(\frac{0.333}{1-0.333}\right)}{\left(\frac{0.667}{1-0.667}\right)} = \frac{0.5}{2.0} = \frac{1}{4},$$

is known as the *odds ratio* (OR), which facilitates calculation of the coefficients β_0 and β_1 of logistic regression in (3.12). Gelman and Hill [75, p. 82] note that an advantage of using ORs over probabilities is that they can be scaled repeatedly without ever reaching either of the boundaries 0 or 1.

The result of taking logarithms on both sides of (3.13) is

$$\log\left(\frac{P(Y_i \mid X_i)}{1 - P(Y_i \mid X_i)}\right) = \beta_0 + \beta_1 X_i, \quad i \in \mathcal{N}, \tag{3.14}$$

in which the left hand side is known as *logit* or *log-odds* [99, p. 132].

3.4.2 Regression coefficient estimation by fitting logistic regression models

Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of the coefficients β_0 and β_1 in (3.14), respectively, such that when they are substituted into (3.12), the results correspond as closely as possible to a set of observed data points, but not too closely so as to avoid *over-fitting* the data set used to build the model [99, p. 132].

Shaliziz [199, p. 227] states that because logistic regression predicts probabilities, as opposed to only class membership, a logistic regression model can be fitted using a likelihood function. Hastie *et al.* [88, p. 120] concur and add that the method of *maximum likelihood* can be used to fit the logistic function and estimate the coefficients. James *et al.* [99, p. 132] acknowledge that it would also be possible to use the method of non-linear least squares to fit the model, but agree that it is generally more acceptable to use the method of maximum likelihood, and probably best due to its statistical properties that are better than those of the other options.

James *et al.* [99, p. 132] summarise the idea behind using maximum likelihood to fit a logistic regression model as follows. The aim is to estimate β_1 and β_0 in (3.14), *i.e.* to find $\hat{\beta}_1$ and $\hat{\beta}_0$, such that when they are substituted into (3.12), the results correspond as closely as possible to the observed points in the data set used to build the model.

Shaliziz [199, p. 227] adds that for each data point there exists a feature X_i and an observed binary class Y_i. As stated above, for the class Y_i, the probability that $Y_i = 1$ is P. The probability is therefore $1 - P$ that $Y_i = 0$. With $P(X_i)$ denoting the probability of the event of interest occurring, the *likelihood function* is

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} P(Y_i \mid X_i)^{Y_i}(1 - P(Y_i \mid X_i))^{1-Y_i}, \quad i \in \mathcal{N} \tag{3.15}$$

and its logarithm is

$$\begin{aligned}\ell(\beta_0, \beta_1) &= \sum_{i=1}^{n}\big[Y_i \log(P(Y_i \mid X_i)) + (1 - Y_i)\log(1 - P(Y_i \mid X_i))\big] \\ &= \sum_{i=1}^{n} \log(1 - P(Y_i \mid X_i)) + \sum_{i=1}^{n} Y_i \log\left(\frac{P(Y_i \mid X_i)}{1 - P(Y_i \mid X_i)}\right), \quad i \in \mathcal{N}. \end{aligned} \tag{3.16}$$

From (3.12) and (3.13) it is possible to obtain the expression

$$1 - P(Y_i \mid X_i) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_i}}, \quad i \in \mathcal{N}, \tag{3.17}$$

which, along with (3.14), may be substituted into the right-hand side of (3.16) to obtain the likelihood function

$$\begin{aligned} \ell(\beta_0, \beta_1) &= \sum_{i=1}^{n} \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_i}}\right) + \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) \\ &= \sum_{i=1}^{n} -\log(1 + e^{\beta_0 + \beta_1 X_i}) + \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i), \quad i \in \mathcal{N} \end{aligned} \tag{3.18}$$

[115, p. 564]. The next step is to find the coefficient estimates by maximising the likelihood function (3.18). This is achieved by differentiating the likelihood function with respect to each coefficient, in this case β_0 and β_1, and equating the results to zero. According to Kutner *et al.* [115, p. 564], one will not, however, be able to solve these equations exactly since there exists no closed-form solution for the values of β_0 and β_1 in the likelihood function (3.18). It is, however, possible to approximate an optimal solution numerically. Two of the most popular numerical methods employed for this purpose are the *Newton-Raphson algorithm* and *Fisher's Scoring* [88, p. 120], [118, p. 1].

The next step, according to Kutner *et al.* [115, p. 564], is to substitute the coefficient estimates into (3.11) or (3.12) so as to obtain the probability of the event of interest occurring.

3.4.3 Selecting a cut-off point

According to Luna [127, p. 32], the probability values between 0 and 1 produced from (3.11) or (3.12) have to be sorted into two classes in order to classify or predict a binary response of an event occurring. This is performed by determining a "cut-off" point or threshold value. If the probability produced by the model is equal to or greater than the predetermined "cut-off" point, the observation is assigned 1. If, however, the probability falls below the "cut-off" value, the observation is assigned 0.

By raising or lowering the "cut-off" point it is possible, to some extent, to increase or decrease the number of correct positive predictions and correct negative prediction. A typical "cut-off" used is 0.5 (*i.e.* observations with a predicted probability of 50% or greater of occurring are assigned the label indicating that the event will occur, and the opposite for those below 50%) [170, p. 160], [175, p. 8], [207, p. 1416].

3.5 Evaluation of a logistic regression model

Besides considering whether or not the assumptions of logistic regression hold (a topic to be discussed later in this thesis), there are four stages in the process of evaluating a logistic regression model, namely evaluation of the overall model, determining the statistical significance of individual coefficients, ascertaining the predictive ability of the model, and validation of the final model [170, p. 158], [175, p. 5]. These four stages are discussed in some detail in this section.

3.5.1 Evaluation of the overall model

The evaluation of the overall model refers to the step of investigating the strength of the relationship between the dependent variable and the combined independent variables investigated. This is performed by comparing the null model (or intercept only model) which contains no independent variables with the full model containing all the independent variables [175].

The null hypothesis and alternative hypothesis used to evaluate an overall model with m independent variables is

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0, \tag{3.19}$$

and

$$H_a : (\beta_1 \neq 0) \text{ and/or } (\beta_2 \neq 0) \text{ and/or } \cdots \text{ and/or } (\beta_m \neq 0), \tag{3.20}$$

respectively. Statistical tests which may be performed to evaluate H_0 include the *likelihood ratio test*, the *Chi-Squared Goodness of Fit test*, and the *Hosmer-Lemeshow test*. The popular likelihood ratio test is discussed here in further detail.

The log likelihood ratio test statistic may be to used compare the deviations of the null model from the full model [94, p. 12]. This allows for an interpretation of the effect that the combined set of independent variables has on the dependent variable. According to Park [170, p. 158], this goodness of fit index is

$$G = -2 \log \frac{\text{likelihood of null model}}{\text{likelihood of full model}}. \tag{3.21}$$

In the case of a full model having m independent variables, the likelihood ratio test follows a χ^2-distribution with z degrees of freedom [91, p. 248].

If (3.21) produces a positive value with a p-value less than the significance level α, the null hypothesis in (3.19) may be rejected and the conclusion may be drawn that at least one of the m independent variables assists in predicting the dependent variable [170, p. 159].

3.5.2 Statistical significance of individual coefficients

If the overall model seems to be a good fit according to one of the statistical tests mentioned above, the next step is to evaluate each of the independent variables separately. Methods which may be used to evaluate the significance of each independent variable include the likelihood ratio test (the same as above, except that only one independent variable is present in the full model), *Wald statistic*, p-value, $\hat{\beta}$-coefficient CI, and the ORs [170, p. 158], [175, p. 5]. Use of the Wald statistic, p-value, $\hat{\beta}$-coefficient CI, and the ORs for this purpose is discussed in more detail in this section.

The Wald statistic, p-value and $\hat{\beta}$-coefficient CI

According to Bewick *et al.* [20, p. 114], the test statistic known as the Wald statistic follows a χ^2-distribution, and is calculated for a particular model variable β_i as

$$W = \left(\frac{\hat{\beta}_i}{\text{SE}_i} \right)^2, \tag{3.22}$$

where $\hat{\beta}_i$ again refers to the estimated value of β_i and SE_i denotes the standard error estimation of that coefficient. The original form of the numerator of (3.22) is $\hat{\beta}_i - \beta_{i,\text{null}}$, where $\beta_{i,\text{null}}$ denotes the value of the parameter of interest in the null hypothesis such that $H_0 : \beta_{i,\text{null}}$, but since the $\beta_{i,\text{null}}$-value of this study is zero, it falls away. A variable is considered statistically significant if $W > c$, where c is a pre-determined critical value. The critical value is typically chosen as $F^{-1}(1-\alpha)$, where F is the cumulative distribution function of a χ^2 random variable with one degree of freedom and a significance level of α [176].

A p-value may also be derived from a z-score, which is derived from $\hat{\beta}_i/\text{SE}_i$, to be compared against α. Using either of the above two methods it is then possible to reject or not reject a null hypothesis stating whether or not the single independent variable x_i is statistically significant.

In addition to the p-value or Wald statistic, that indicate statistical significance, the CIs around the $\hat{\beta}_i$ may be utilised to investigate the results further, as explained in §3.1.3. The CI of the $\hat{\beta}_i$, may be found by

$$(100-\alpha)\% \text{ CI for } \hat{\beta}_i = \hat{\beta}_i \pm z_{1-\alpha/2}\text{SE}_i,$$

where $\hat{\beta}_i$, $z_{1-\alpha/2}$, and SE_i denote the coefficient of the estimate, the $100(1-\alpha/2)$ percentile z-score, and the standard error estimate, respectively [107, p. 376], [196, Chapter 39].

Odds ratios

Recall from §3.4.1 that the left-hand side of (3.13) is known as the odds [75, p. 82]. These odds take the form $\frac{P}{1-P}$ for an event which has a probability P of occurring and a probability $1-P$ of not occurring.

The resulting OR value falls in one of the following three categories [221, p. 227]:

- OR $= 1$ suggests that the independent variable under consideration has no effect on the dependent variable,
- OR > 1 suggests that an increase in the exposure of the independent variable in question would result in an increased odds of the desired outcome involving the dependent variable, and
- OR < 1 suggests that an increase in the exposure of the independent variable in question would result in a lower odds of the desired outcome involving the dependent variable.

The precision of the OR may be validated by means of a CI, as explained in detail in §3.1.3. The CI may be utilised as a proxy for the presence of statistical significance, provided that the interval does not overlap with the null value. The null value for the OR is 1 [221, p. 227]. According to Park [170, p. 160], the limits of the CI for an OR_i may be calculated as

$$(100-\alpha)\% \text{ CI for } \text{OR}_i = e^{\ln \text{OR}_i \pm z_{1-\alpha/2}(\text{SE}_i[\ln \text{OR}_i])}, \tag{3.23}$$

where $\ln \text{OR}_i$ is the log of the OR_i and $\text{SE}_i[\ln \text{OR}_i]$ denotes the standard error of the log of the OR_i. James *et al.* [99, p. 67, p. 132] state that a small CI-value suggests a high precision of the OR, while a high CI-value suggests a low precision of the OR, but as was mentioned in §3.1.3 both a small p-value and narrow CI may simply be the result of a large sample.

Table 3.2 contains examples of ORs and 95% CI limits for three logistic regression variables in the context of the hypotheses

$$H_0 : \text{OR} = 1.0$$

and

$$H_a : \text{OR} \neq 1.0.$$

By studying Table 3.2, it may be stated that a one-unit increase in *Variable A* results in an increase in the odds of the desired event occurring (*versus* not occurring) by a factor of 1.824. Similarly, the odds of the desired event occurring increases by a factor of 0.745 (thus decreases) for each unit increase of *Variable B*. Notice, however, that although *Variable C* has a reasonably high OR, its CI overlaps with the variable of interest specified by H_0 of 1.0, implying that it is not statistically significant [99, p. 133].

Table 3.2: *Example of OR and 95% CI boundaries.*

Independent variable	OR	2.5% CI	97.5% CI
Variable A	1.824	1.126	1.978
Variable B	0.745	0.224	0.812
Variable C	1.745	0.941	2.241

It is also necessary to correctly interpret the ORs in the case that dummy variables and a reference class is used. Consider again the example presented in (3.10), in which the *Race* category of *Black* is used as the reference class and the remaining *Race* categories of *Coloured*, *Indian*, and *White* are left in. Suppose the OR-values of Table 3.3 are the results produced for this example. Keeping in mind that the race *Black* is the reference class, according to Lin and Fultz [122], the correct interpretation of these results are as follows: The OR-value of 5.451 for race *Indian* is the odds of an *Indian* student being successful at tertiary level divided by the odds of a *Black* student being successful. The two remaining categories may be interpreted in the same fashion.

Table 3.3: *Example OR results for* (3.10).

Independent variable	OR
Coloured	0.820
Indian	5.451
White	3.215

3.5.3 Predictive ability of the model

The third part of logistic regression model evaluation involves an assessment of how well the model can predict dependent variable outcomes based on a set of given independent variables.

For each observation of a sample, the logistic regression function (3.12) produces a probability value between zero and one. Depending on which side of the "cut-off" threshold the calculated probability falls, the observation will be assigned a label of either 1 or 0. The standard "cut-off" used is 0.5.

When comparing the predicted and actual outcome of specific observations, only four possibilities exist, as shown in Table 3.4 [169, p. 46].

An observation is considered to fall within the true positive category if it was predicted that the observation would be positive and the actual outcome is also positive. Similarly, the true

Table 3.4: *Classification table for comparing predictive outcomes versus actual outcomes.*

		Actual outcome:	
		Positive	Negative
Predicted outcome:	Positive	(a) True positive	(b) False positive
	Negative	(c) False negative	(d) True negative

negative category refers to observations which were predicted to be negative and was indeed actually negative. Observations may also be predicted incorrectly. If the observation is predicted to be positive and is, in fact, negative, then the observation is categorised as false positive (a type I error). In the same manner, if an observation is predicted to be negative, but is, in fact, actually positive, then it is a false negative (a type II error).

By counting how many observations fall in each of the four classes mentioned above, Altman and Bland [8] state that it is possible to calculate two useful statistics known as *sensitivity* and *specificity*, which may be employed to determine the proportions of binary outcomes that have been predicted correctly. The total count of each of the four classes may be represented by a (true positive), b (false positive), c (false positive), and d (true negative). Sensitivity, also known as the *true positive rate*, refers to the proportion of positive predictions made correctly, while specificity, also known as the *true negative rate*, refers to the proportion of negative predictions made correctly. Sensitivity may therefore be calculated as

$$\frac{a}{a+c}, \tag{3.24}$$

and specificity may be calculated as

$$\frac{d}{b+d} \tag{3.25}$$

[169, p. 46].

3.5.4 Validation of the final model

Validation of the final model, the fourth step identified by Park [170, p. 160], simply refers to assessing the predictive performance of a model, as discussed in the previous step (§3.5.3), but by using previously unseen data. Exactly how this may be achieved has previously been discussed in §3.1.10 and §3.1.11.

3.6 Classification and regression trees

The tree-based algorithm known as *classification and regression trees* (CARTs) is a popular binary recursive partitioning method for both classification and regression [138, p. 145]. Luna [127, p. 8] explains that CART is a non-parametric method, as described in §3.1.5. The method is considered *binary* as each parent node in a tree-like data structure is partitioned into two child nodes. The term *recursive* refers to how each new child node may, in turn, be treated as a parent node so as to allow for the process to repeat itself. *Partitioning* refers to the fact that the method separates or partitions the data into a number of subsets. According to Yohannes and Hoddinott [252, p. 1], a *regression tree* is produced if the dependent variable is continuous while the result is a *classification tree* if the dependent variable is categorical. For both regression and classification trees the independent variables may be either numerical or categorical [84].

Since the CART method is not based on a probabilistic model, the predictions and variable importance results of individual variables cannot be associated with CIs or probability levels. Confidence in the accuracy (*i.e.* a measure of the misclassification rate) of the results produced by a CART analysis is thus based on the historical accuracy of the tree [252, p. 11].

The two main steps in a CART analysis are: (1) Tree growing and node splitting, and (2) computing the optimal tree size. A discussion on these steps follows in the remainder of this section, with a focus on classification trees, since this is the paradigm in which the algorithm is applied later in this thesis.

3.6.1 Tree growing and node splitting

A classification tree is produced by growing a tree, starting with a trunk and generating branches from this trunk which form new nodes. Each node contains a specific subset of the entire sample. The trunk in this instance is the first node, also called the *root node*, and contains the entire sample. During a CART analysis many *if-then* split conditions are generated which allow for a classification of data cases [212]. Starting with the first node, the sample is partitioned so as to maximise a certain splitting criterion. This process is repeated for each new node until a stopping criterion is satisfied or there exists a node for each case in the data. *Terminal nodes* are the bottom nodes or endpoints in the CART tree which represent the final classification outcomes in the tree [84].

In order to explain the tree growing and splitting process mathematically, consider a training sample matrix $\boldsymbol{X}$ with m dimensions (the total number of independent variables) and n cases (the total number of observations). Thus there are m variables $x_1, x_2, \ldots, x_m$, and suppose the optimal splitting value for variable x_j is denoted by x'_j. All instances of variable x_j with values less than x'_j are partitioned to the left child node and the remaining instances to the right child node of the current node [185, p. 75].

Consider t_p as the parent node and its left and right child nodes as t_ℓ and t_r respectively, as visualised in Figure 3.18.

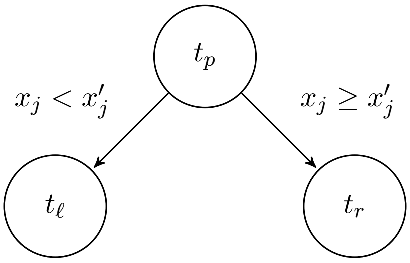


FIGURE 3.18: *Parent and child nodes in a CART.*

According to Lewis [119, p. 6], the general aim of splitting at each node is to achieve the largest improvement in predictive accuracy, which may be achieved by maximising the *purity* or *homogeneity* of the two subsequent child nodes. This improvement in accuracy may be achieved by employing a measure known as the *splitting criterion*, *splitting function*, *impurity measure* or *impurity function* [212], [226, p. 10].

At each split point (parent node), denoted by t, the *impurity function*, denoted by $i(t)$, depends on the impurity of the parent node as well as those of both child nodes. The impurity of the parent node remains constant regardless of the split and is denoted by $i(t_p)$. If the combined impurity of the two child nodes is denoted by $i(t_c)$, and the expected impurity of the two child nodes is denoted by $E[i(t_c)]$, the change in impurity of the parent and child nodes may be written as

$$\Delta i(t) = i(t_p) - E[i(t_c)]. \tag{3.26}$$

By letting the probabilities of the left and right node classifications be P_ℓ and P_r, respectively, (3.26) becomes

$$\Delta i(t) = i(t_p) - P_\ell i(t_\ell) - P_r i(t_r).$$

For classification trees it is required that all possible values of all variables in the matrix $\boldsymbol{X}$ must be considered in search of the best split, $x_j < x_j'$, $j = 1, 2, \ldots, m$, which results in the change in impurity function $\Delta i(t)$ being maximised. This process may be expressed mathematically[1]at each parent node as

$$\underset{x_j < x_j',\, j=1,2,\ldots,m}{\operatorname{argmax}} \quad [i(t_p) - P_\ell i(t_\ell) - P_r i(t_r)]. \tag{3.27}$$

By maximising the change in impurity function, the purity and relative homogeneity of the two child nodes are maximised [119, p. 6], [226, p. 10]. Although the process of maximising the change in impurity function (3.27) applies to both classification and regression trees, the impurity function $i(t)$ is defined differently for each of these cases [152, p. 583].

For regression trees two popular optimisation methods are the method of *least squares* and the method of *least absolute deviations*. In the method of least squares, the aim at each node is to minimise the sum of the squared errors between the mean in each node and the observation. In the method of least absolute deviations, on the other hand, the mean absolute deviation from the median within a node is minimised.

Although other criteria exist, the four main criteria used for impurity minimisation, according to Moisen [152, p. 584], are the *Gini index*, the *misclassification error*, the *entropy index*, and the process of *twoing*.

According to Timofeev [226, p. 10], the *Gini splitting rule* or *Gini index* is the most popular impurity function for classification trees. Let the total number of classes be K. Also, let the proportion of observations of class k in node t be p_{tk}. Then the Gini splitting rule is to minimise $\sum_{k=1}^{K} p_{tk}(1 - p_{tk})$.

The misclassification error measure determines the proportion of observations that are not part of the majority class for each node [152, p. 584].

The entropy index, also known as the *deviance* or *cross-entropy measure* of impurity, is given by $\sum_{k=1}^{K} p_{tk} \log p_{tk}$. This function has to be minimised. Both the Gini index and the entropy index are more sensitive than the misclassification error to changes in the node probability [152, p. 584].

The process of twoing was designed specifically for multiclass problems, but approaches each multiclass problem as a binary problem. The method favours those splits that allow for related classes to remain together and thus searches for two classes which, when combined, will encompass more than half of the data [226, p. 11].

[1] While the traditional operation max $f(x)$ would provide the maximum value of $f(x)$ obtainable by varying x, the operation argmax $f(x)$ provides the value of x at which this maximum is obtained.

3.6.2 The optimal tree size

An overgrown tree, which fits the data almost exactly, will result in overfitting and exhibit poor predictive ability in respect of an independent test set. On the other hand, a tree that is too small does not contain enough tree splits to discover meaningful relationships in the data. The ideal is therefore a tree of a suitable trade-off size. Moisen [152, p. 585], explains that a tree may be grown to its optimal size either by employing a *stopping criterion* or by employing a process of *pruning*.

The use of stopping criteria, as the name implies, forces the tree growing process to stop once a certain criterion is met. Two examples of such stopping criteria are described by Rashidi *et al.* [185, p. 76]. These are the *minimum node size to split* criterion according to which the tree growing process terminates at a node if the number of observations at that node is smaller than a specified minimum number [226, p. 15], and the *maximum tree depth* criterion according to which a maximum number of split levels that may be present in a tree is pre-specified.

The more popular option, as opposed to the use of a stopping criterion, is to adopt some method of pruning. Pruning requires that a tree is grown to full size using all the available data, and such a tree is called a *maximum tree*. Such a maximum tree is then pruned back to a suitable size. The optimal size to which a decision tree should be pruned back may be determined by any of a number of methods. These include the method of *reduced-error pruning*, *pessimistic pruning*, *rule post-pruning*, and *cost complexity pruning* [99, p. 308], [189, p. 70].

Several studies have been done to compare the performance of various pruning methods against each other, including studies by Esposito *et al.* [64], Mingers [149], and Quinlan [183]. The majority of these studies found that no single pruning method performed better than all others under all circumstances. Due to this finding it was decided that the method of cost complexity pruning, referred to by some in the literature as *weakest link pruning* or *error complexity pruning*, will be sufficient for the purposes of this thesis due to its popularity in the literature. Since the method of cost complexity pruning is later applied in this thesis, it is described in the remainder of this section [99, p. 308], [189, p. 70].

According to Lewis [119, p. 8], the method of cost complexity pruning attempts to find an optimal balance between a tree's complexity (the number of terminal nodes in the tree) and its misclassification rate (also known as its *prediction accuracy* or *average decision cost*). For the test data, the misclassification rate of a tree continuously decreases as the complexity of the tree increases up to some point, after which it will start to increase. The aim is thus to find the point at which the misclassification rate achieved by the tree in respect of the test data is at its lowest. The maximum tree may then be pruned back to that point so as to reveal the optimal tree. This process is illustrated graphically in Figure 3.19.

Denote the misclassification rate or *re-substitution error* of a tree T by $R(T)$. This misclassification rate indicates the proportion of observations in a training sample which are misclassified by T. Also, let $|\tilde{T}|$ be the number of terminal nodes in T. Then $\lambda|\tilde{T}|$ is a measure of the *complexity* of T, where λ is a real number called the *complexity parameter*. The *cost complexity function* of T is defined as

$$C_\lambda(T) = R(T) + \lambda|\tilde{T}| \tag{3.28}$$

[226, p. 17]. Assume now that $T(\lambda)$ represents the smallest subtree pruned from $T_{\max}$ with the same misclassification rate $R(T)$ as that of T for a fixed value of the complexity parameter λ.

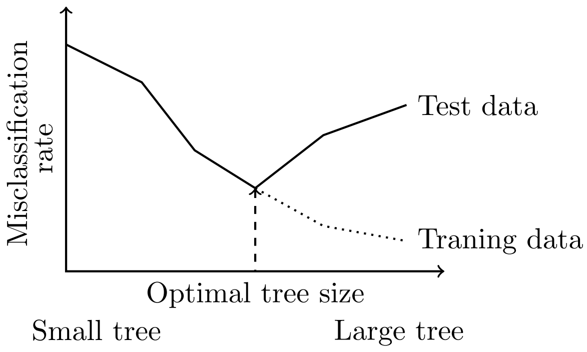


FIGURE 3.19: *Misclassification rate versus tree size.*

Such a tree, $T(\lambda)$, will satisfy the condition

$$C_\lambda(T(\lambda)) = \min_{T \subseteq T_{\max}} C_\lambda(T) \tag{3.29}$$

and if

$$C_\lambda(T) = C_\lambda(T(\lambda)), \tag{3.30}$$

then

$$T(\lambda) \subseteq T. \tag{3.31}$$

The minimisation process in (3.29) implies that there exists no subtree of $T_{\max}$ which has a lower cost complexity than the tree $T(\lambda)$, while (3.30) and (3.31) together state that in the event of more than one tree having the same minimum cost, the smaller tree is chosen.

For every value of $\lambda \in [\lambda_k, \lambda_{k+1})$, there exists a smallest minimising subtree T_k for each natural number k. Note that there exists an infinite number of permissible λ values, but only a finite number of subtrees.

If T_1 is the first smallest minimising subtree of $T_{\max}$ (*i.e.* $k = 1$) found for $\lambda = 0$, then it can be shown that the smallest minimising subtrees to follow satisfy

$$T_k \supset T_{k+1} \supset T_{k+2} \supset \cdots \supset T_{\text{root}}, \tag{3.32}$$

where T_{root} represents the root node [177]. The nested relationship (3.32) implies that T_{k+1} may be found by pruning T_k, that T_{k+2} may be found by pruning T_{k+1}, and so on. By this logic it is possible to find the next optimal tree in the sequence (3.32) by increasing λ, rather than having to start from the maximum tree again. The process starts with $\lambda = 0$, resulting in the largest tree selected as the penalty is zeroed, but the tree will become only the root node as λ approaches infinity.

Algorithm 3.1 may be used to compute the first tree T_1 in the sequence from the maximisation tree $T_{\max}$. The goal is to find the smallest subtree of $T_{\max}$, while keeping $\lambda = 0$, that has the same misclassification rate as $T_{\max}$. In the algorithm, I is the total number of parent nodes in a tree, t_p^i is the parent node i, and t_ℓ^i and t_r^i are its left and right child nodes, respectively.

For each tree in (3.32) there exist certain nodes which, when pruned away, results in a subtree that has equal cost complexity to the original, but since it is smaller it is considered better than the original. These nodes, sometimes referred to as the *weakest links*, may be found and removed in the following manner.

Algorithm 3.1: Determining the tree T_1

Input : The maximum tree, $T_{\max}$.
Output: The first cost minimisation tree, T_1.

```
T' ← T_max
while there exists a parent child relationships of $R(t_p) = R(t_\ell) + R(t_r)$ do
    for $i \leftarrow 1$ to $I$ do
        if $R(t_p^i) = R(t_\ell^i) + R(t_r^i)$ then
            $T' \leftarrow T' - T_{t_p^i}$
$T_1 \leftarrow T'$
```

Denote a branch rooted at node t by T_t. The cost complexity function in (3.28) may be defined for a specific node and its branch as

$$C_\lambda(t) = R(t) + \lambda, \tag{3.33}$$

and

$$C_\lambda(T_t) = R(T_t) + \lambda|\tilde{T}_t|, \tag{3.34}$$

respectively. It is known, for a sufficiently small value of λ, that the inequality $C_\lambda(T_t) < C_\lambda(t)$ holds. Also, as λ increases, $C_\lambda(T_t)$ will increase faster than $C_\lambda(t)$ since λ has a larger positive effect on the function $C_\lambda(T_t)$. As λ increases, a point will be reached where $C_\lambda(T_t) = C_\lambda(t)$, after which $C_\lambda(T_t) > C_\lambda(t)$. The λ-value for which the two functions (3.33) and (3.34) are equal is denoted by λ_{cut}, as depicted in Figure 3.20. Although λ_{cut} indicates the point where T_k and $T_k - T_t$ have the same cost complexity, $T_k - T_t$ is the smaller of the two trees, which is thus considered better.

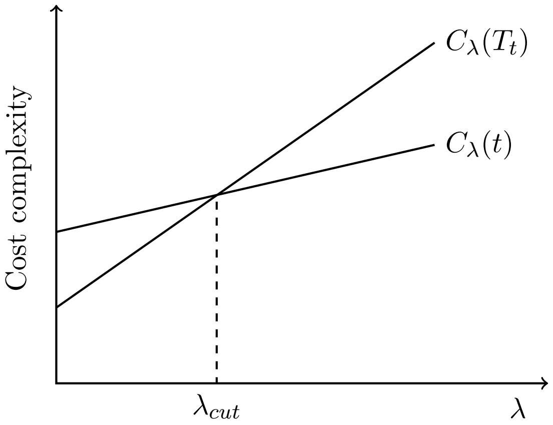


FIGURE 3.20: *Cost complexity of a single node t and its branch T_t as λ increases.*

If a tree T_k is pruned at node t and the branch being pruned is denoted by T_{t_k}, then the misclassification cost will increase by $R(t)-R(T_{t_k})$ (note that the increase depends on the branch being pruned, not on the random original tree) and its complexity (the number of terminal nodes in the resulting tree) decreases by $|\tilde{T}_{t_k}|-1$. These quantities may be combined to form the ratio

$$g(t_k) = \frac{R(t) - R(T_{t_k})}{|\tilde{T}_{t_k}| - 1}, \tag{3.35}$$

which represents the λ_{cut}-value for each node t_k of the tree T_k and indicates the ratio of increased misclassification cost per node if pruned. Using (3.35), the ratio for each node of the tree T_k may be determined, and from these the smallest ratio, denoted by $g(\bar{t}_k)$, may be selected, which is the weakest link node of that tree. It may thus be stated that $\lambda_k = g(\bar{t}_k) = \min g(t_k)$. There may be more than one node with the same value of $g(\bar{t}_k)$. These weakest node(s) may be removed from the current tree T_k so as to find the next tree $T_{k+1} = T_k - T_{\bar{t}_k}$ in (3.32).

This new tree is then considered as the new starting tree and the entire process is repeated until the tree T_{root} is reached. The above process is summarised in Algorithm 3.2. Note that the initial tree of the algorithm is not $T_{\max}$, but rather T_1, as determined by means of Algorithm 3.1.

Algorithm 3.2: Computing $T_{k+1} \supset T_{k+2} \supset \cdots \supset T_{\text{root}}$

Input : T_1 determined using Algorithm 3.1.
Output: The sequence of pruned trees $T_{k+1} \supset T_{k+2} \supset \cdots \supset T_{\text{root}}$ in (3.32).

```
$k \leftarrow 1$
while $T_k \supset T_{root}$ do
    for each non-terminal node in $T_k$ do
        $g(t_k) \leftarrow \frac{R(t)-R(T_{t_k})}{|\tilde{T}_{t_k}|-1}$
        $\lambda_k \leftarrow g(\bar{t}_k) = \min g(t_k)$
        $T_{k+1} \leftarrow T_k - T_{\bar{t}_k}$
        $k \leftarrow k+1$
```

3.6.3 The optimal subtree

From the above series of pruned trees, one must be selected as optimal based on a certain minimum error criterion. This may be achieved by use of assigned classes to nodes and through the use of testing the tree in respect of data not selected to be part of a specific bootstrap sample, known as the *out-of-bag* (OOB) observations, in the context of the test set [189].

Outcome classes are assigned to nodes in the following manner. Once the tree has been pruned to optimal size, each of the nodes are assigned a class (in the case of a classification tree) or a response value (in the case of a regression tree). Lewis [119, p. 7] points out that each node, even the root node, is assigned a class. Timofeev [226, p. 18] emphasises, however, that the only nodes which are of importance with respect to having classes assigned to them are the terminal nodes. For each terminal node in a classification tree, the so-called *dominating class* is assigned to it. The dominating class is the class most represented by observations of the training data in a specific terminal node. Thus, when an independent test set of OOB observations is presented to the tree, each observation is assigned to a terminal node, based on the answers to the tree's questions at the previous nodes, and is assigned an outcome class accordingly.

Once the nodes have been assigned classes, the minimum error rate of the tree may be assessed using the test set in conjunction with the technique of K-fold CV, as discussed in §3.1.11. CV does not require an independent test data set. The available data are rather partitioned into K folds of equal size, typically adopting the value $K = 10$. The entire process of tree building and pruning, discussed above, is conducted K times based in the learning data. After each run, the error rates all of the trees in (3.32) are evaluated in respect of the set held out. Once this process has been conducted K times, all equally sized trees (same number of terminal nodes) are grouped together and their error rates averaged for each tree size. The tree size, and subsequent λ-value, associated with the smallest average error rate, is selected as the optimal subtree. In this manner it is possible to evaluate the misclassification rate of trees based on their complexity values [119, p. 9].

3.6.4 CART enhancements

Although CART is a powerful predictive model in its own right, Moisen [152, p. 587] describe three attractive enhancements of CART, namely incorporation of the notions of *bagging*, *boosting*, and *random forests*.

Bagging is modelled on the process of *bootstrap re-sampling*, as discussed in §3.1.11. Starting with a training sample S, B bootstrap samples are drawn from S, yielding the bootstrapped training set $S_1, S_2, \ldots, S_B$. Although bagging may be applied to any single statistical learning technique, the CART model is considered here to illustrate the working of the notion of bagging [99, p. 320]. A separate CART model is trained on each of the bootstrapped training sets. Each CART model's predictions are made in respect of the remaining data not selected to be part of a specific bootstrap sample, *i.e.* the OOB observations. The results of testing in the context of the OOB samples may be averaged (for regression trees) or the majority vote may be determined (for classification trees) so as to obtain final predictions for each observation [99, p. 320]. According to Hastie *et al.* [88, p. 282], these bagging averages exhibit less variance than the predictions produced by a single CART model.

Unlike bagging, which makes use of bootstrapped sample training sets and builds all the trees before evaluating them, *boosting* builds trees sequentially and makes use of an adapted sample as each new tree utilises information from previous trees [99, p. 321]. Initially all observations are given the same weight, but with each new model built, those observations that are regularly misclassified are given more weight, resulting in their more frequent selection in the new training samples of future models. Such observations are known as *problematic observations*. Each new model is therefore built on a re-weighted version of the original training data and in this manner each new model should exhibit better prediction ability [152, p. 587]. James *et al.* [99, p. 321] note that, unlike bagging, boosting may result in overfitting if the number of trees grown is too large. Examples of boosting algorithms include *Adaboost* [42], *BrownBoost* [141], *LogitBoost* [141], and *LPBoost* [194].

The third type of CART enhancement, *random forests*, is discussed as a classification scheme in the its own right in the next section.

3.7 Random forests

The method of *random forests* builds on the notion of bagging (§3.6.4) and utilises many trees produced by CART analyses so as to obtain improved prediction ability. The same approach is followed as explained above for bagging, were each tree is built using a bootstrapped sample drawn from the original sample. Suppose the number of random samples drawn, and thus the number of trees drawn, is denoted by *ntree*. In addition to the randomness induced by the bootstrapped sample, Shi and Song [202, p. 184] explain that one more parameter is defined and applied so as to bring about even more randomness to the trees. During the CART analysis each of the m variables are considered to be a possible splitting variable at each node, but with random forests only a subset of cardinality *mtry* of all the variables is considered as possible splitting variables. The parameter *mtry* is typically taken as $\sqrt{m}$, where m is the total number of independent variables present in the model [99, p. 320].

The process of selecting a bootstrapped sample and splitting in respect of only *mtry* randomly selected variables at each node is repeated *ntree* times. The resulting trees are grown to maximum size or bounded as specified by a minimal node size criterion. Either way, the trees are all left un-pruned. As with the previous ensemble methods, the final predictions in the context of the test observations are either the majority vote (for classification trees) or the average (for regression trees) [152, p. 587].

3.7.1 Variable importance measure using random forests

In addition to using classification models to predict outcomes for a data set, it is also often employed to determine the importance of each independent variable in the context of a specific model [126, p. 1]. Random forests provide this functionality by determining the so-called *variable importance measure* (VIM) for each independent variable. Two popular methods for determining the VIM are the *mean decrease Gini* method, also known as the *Gini importance* method, and the method of *mean decrease accuracy* [126, pp. 2–3].

In a study conducted by Strobl *et al.* [219], it was found that the mean decrease Gini method is biased towards independent variables that have many values or categories. In a second study, Strobl *et al.* [218] claimed that the Mean Decrease Accuracy method is biased in the sense that it overestimates the importance of correlated variables, but in a study conducted two years later by Genuer *et al.* [76] the authors were unable to confirm the claims made by Strobl *et al.* [218] concerning the Mean Decrease Accuracy method. Due to the above developments it was decided that the mean decrease accuracy method will be utilised later in this thesis. This method is therefore discussed in more detail in the remainder of this section.

According to Gislason *et al.* [79, p. 295], the VIM method of mean decrease accuracy functions as follows. For each of the trees produced, the outcomes for the OOB observations are estimated and compared against the actual outcomes of those observations so as to determine the predictive accuracy of the tree. For the same tree the OOB observation values of a single independent variable are permuted, thus breaking the original relations between the dependent variable. As described above, the predictive accuracy of the tree is assessed and the decrease of predictive accuracy of the tree is noted. In the case where the independent variable permuted is significant, the accuracy will noticeably decrease. This is repeated for each independent variable in each tree of the method of random forests, and the variable importance of each independent variable is averaged across all the trees so as to produce a final VIM for that independent variable [121, p. 18]. According to Ishwaran [98, p. 522], a negative or zero VIM value signifies an independent variable that does not contribute positively to the process of prediction. Since permutation is employed in this method, mean decrease accuracy is also known as the *permutation accuracy importance* or *permutation importance* [126, p. 3].

3.7.2 Outliers for random forests using proximity measures

Random forests incorporate a measure of the proximity between observations, by use of a so-called *proximity matrix*. According to Gislason *et al.* [79, p. 296], the composition of the proximity values matrix works as follows. In a sample of n observations, consider each pair of observations i and j. For each terminal node of all the random trees created in which the pair i and j are classified together in a terminal node, their proximity, denoted by $\text{prox}(i, j)$, is increased by one. Each $\text{prox}(i, j)$-value may then be normalised by dividing it by the number of trees created. The results of the above is an $n \times n$ proximity matrix indicating in what fraction of trees observations i and j belong to the same terminal node [121, p. 18].

The proximity matrix may be evaluated further so as to detect outliers in the data in the following manner. For each observation i, the average squared proximity is calculated for all pairs $\text{prox}(i, j)$ in which i and j belong to the same class. The result is known as the *average squared proximity* of observation i with respect to all the other observations of the same class. Thereafter, the *outlier measure* for observation i is determined, which is the number of observations belonging to the same class divided by the average squared proximity of i. For each class the median and absolute deviation are determined for the outlier measure values. Finally, the individual outlier measure value of each observation i is normalised by subtracting the class median from its value and dividing the result by the absolute deviation of the class.

In summary, the process described above simply assigns a score to each observation based on its distance from the centroid of its outcome class. The resulting value for each observation is called its *outlier score*. Those scores less than zero are set to zero. It is generally accepted in the literature that an observation for which this value exceeds 10 may be considered an outlier [32, p. 37], [79, p. 296].

3.8 The C4.5 algorithm

The C4.5 tree-construction algorithm developed by Quinlan [184] is similar to CART (see §3.6) in the sense that it starts with all the training data in a single root node, which is partitioned into subsequent nodes based on a particular splitting criterion. In the case of the C4.5 algorithm the default spitting criterion is the *information gain ratio*.

Ruggieri [192, p. 438] describes the variable on which a node is partitioned by the C4.5 algorithm as follows. In the case that splitting is performed on a quantitative independent variable, the partitioning point may be defined as the *threshold* at which the two subsequent child nodes consist of $t_\ell \leq \textit{threshold}$ and $t_r > \textit{threshold}$, just as in a CART. For qualitative independent variables, however, the C4.5 algorithm produces as many child nodes as there are categories present in the variable (unlike CART). For this reason, the C4.5 algorithm is known as a *non-binary split* or *multi-way split* tree growing algorithm [105, p. 599].

3.8.1 Tree growing in the C4.5 algorithm

The tree growing approach utilised by the C4.5 algorithm, the pseudo code of which is shown in Algorithm 3.3, is a greedy approach focussed on finding the best local choice and does not permit backtracking [71, p. 1], [192, p. 438].
In Step 1, the root node containing all the data is taken as the first parent node t_p. The process from Step 2 onwards is then repeated for the current parent node t_p and each new parent node that is subsequently chosen.
For t_p, the *class frequency* is calculated for both binary response classes in Step 3. The class frequency is the fraction of observations in t_p that belong to each of the binary outcome classes. In Step 4, it is determined whether either of the two classes has a similar member or is greater than a pre-determined cut-off value, in which case t_p is classified a terminal node and is therefore not partitioned further, rather assigning the majority class to t_p. The final part of the if-statement, if its condition found to be true, is to identify a new node t_p if one is available and to repeat the process from Step 3 onwards. Otherwise, the algorithm is terminated.

If the condition of the if-statement of Step 4 is not satisfied, the algorithm continues from Step 7. The information gain is calculated in Step 8 for each independent variable. The best independent variable is identified in terms of information gain and a threshold is identified if the best variable is quantitative.

The operations from Step 11 onward are applicable to each of the c child nodes produced when t_p is partitioned. If the best independent variable is quantitative, then $c = 2$; else $c = h$, where h is the number of categories present in the best qualitative independent variable.

If t_s is empty (Step 12), then it becomes a terminal node, is given the majority class of t_p and is assigned a classification error of 0. If the condition of the if-statement of Step 12 is false, however, t_s becomes the new t_p and the algorithm is repeated from Step 3.

Algorithm 3.3: C4.5 tree construction algorithm.

```
Input   : Training sample & cut-off value.
Output: C4.5 decision tree.
1  Current parent node ($t_p$) ← $T_{\text{root}}$
2  for $t_p$ and each subsequent new $t_p$ do
3      Compute the Class frequency for $t_p$.
4      if Class Frequency of largest class $= 1$ or $> (1 -$ set cut-off$)$ then
5          Convert $t_p$ into a terminal node.
6          Select new $t_p$ if one is available and repeat from Step 3.
7      else
8          for each independent variable and its possible subsets do
9              Compute information gain.
10             Update best information gain independent variable and threshold if best
               splitting variable is quantitative.
11         for each $t_s$ produced by splitting $t_p$ do
12             if $t_s$ is empty then
13                 $t_s$ becomes terminal node.
14                 $t_s$ given majority class of $t_p$ and a classification error of 0.
15             else
16                 $t_p \leftarrow t_s$
17                 Repeat from Step 3.
18     Determine majority class of each node.
```

The final step of the algorithm (Step 18) is to determine the majority class of each node, starting with the terminal nodes. The majority classes of the subsequent parent nodes are calculated as the sums of the misclassification rates (weighted based on number of observations per node) of their direct child nodes.

3.8.2 The C4.5 splitting criterion

The splinting criterion described in this section is unique to the C4.5 algorithm. As mentioned above, the C4.5 algorithm utilises the *information gain ratio* as splitting criterion, which builds on the celebrated *gain criterion* [184, p. 20].

As already stated above, once a node has been identified for splitting it becomes a parent node and produces child nodes $t_1, t_2, \ldots, t_c$. The value of c here depends on the type of independent variable identified for splitting. Denote the number of observations of a subset S of the sample data that belong to class j by $\text{freq}(C_j, S)$ [184, p. 20].

The theory behind the gain criterion is that the total amount of information that is transmitted by a certain message directly depends on the probability of the message, where the message refers to a specific outcome. It may then, for example, be said, if there exists six equally probable messages or outcomes, that the probability of a single message is $\frac{1}{6}$. If the J possible classes are then considered to be the messages or outcomes, it may be said that class j has the probability

$$\frac{\text{freq}(C_j, S)}{|S|} \tag{3.36}$$

of occurring. The information gain is, however, measured in bits. The number of bits of the quantity in (3.36) may be calculated using

$$-\log_2(\text{Probability of outcome}). \tag{3.37}$$

It is possible to combine (3.36) and (3.37) to form

$$-\log_2\left(\frac{\text{freq}(C_j, S)}{|S|}\right),$$

which may be used to determined the message probability in bits [184, p. 21].

The expected information gain associated with a particular subset S of data is determined by summing the quantities above over the outcome classes in proportion to their frequencies within S, to form the measure

$$\text{info}(S) = -\sum_{j=1}^{J} \frac{\text{freq}(C_j, S)}{|S|} \times \log_2\left(\frac{\text{freq}(C_j, S)}{|S|}\right). \tag{3.38}$$

It is possible to apply (3.38) to a specific subset or training set in order to identify the average amount of information required to classify an observation of S into one of the J classes. The quantity in (3.38) is known as the entropy of S and the expression in (3.38) is also known as the *entropy function* [192, p. 439].

If S is assumed to be contained within a single parent node t_p and has to be partitioned based on a specific independent variable X, and thus a parent node t_p, the subsequent child node may then be defined as $t_1, t_2, \ldots, t_c$. The expected information gain for t_p and its child nodes may be found as the weighted sum over child nodes, that is

$$\text{info}_X(t_p) = \sum_{i=1}^{c} \frac{|t_i|}{|t_p|} \times \text{info}(t_i). \tag{3.39}$$

It is finally possible, by subtracting (3.39) from (3.38), to determine the information that is obtained by partitioning a sample S (in parent node t_p) based on a selected independent variable X as

$$\text{gain}(X) = \text{info}(t_p) - \text{info}_X(t_p). \tag{3.40}$$

The information gain criterion (3.40) is utilised by the so-called ID3 algorithm, the predecessor of the C4.5 algorithm. Quinlan [184, p. 24] subsequently built on the information gain criterion to create the so-called *information gain ratio*, which is more robust than the gain criterion and generally produces better results. Using the same fundamentals as in (3.38), a new function is derived which does not, like (3.38), consider the J possible outcome classes, but rather the c possible outcomes of an independent variable X. According to Ruggieri [192, p. 439], this function is defined as

$$\text{split info}(X) = -\sum_{i=1}^{c} \frac{|t_i|}{|t_p|} \times \log_2\left(\frac{|t_i|}{|t_p|}\right). \tag{3.41}$$

Whereas (3.38) determines the information related to partitioning t_p into J classes, (3.41) determines the information created by partitioning t_p into c child nodes based on X.

Finally, it is possible, through the use of the *gain ratio*, to determine the proportion of information generated by partitioning t_p on X by combining (3.40) and (3.41) to form the *gain ratio*

$$\text{gain ratio}(X) = \frac{\text{gain}(X)}{\text{split info}(X)},$$

which is computed in Step 8 of Algorithm 3.3 [184, p. 23].

3.8.3 Pruning the tree in the C4.5 algorithm

As in the pruning process of CART, the C4.5 algorithm reduces the full tree down to a smaller one so as to reduce the effect of overfitting. Quinlan [184, p. 37] explains that although the process of growing overfitted trees and then pruning them back is a more time-consuming process than only growing trees to a predefined set size, the pruned-back trees are more reliable.

In addition to *cost-complexity* pruning, the process employed in a CART analysis which attempts to identify a specific weighting, the C4.5 algorithm employs *reduced-error* pruning which determines the error or misclassification rates directly using the test set. As in *cost-complexity* pruning, the logic of *reduced-error* pruning is as follows [71, p. 1], [104, p. 540], [184, p. 40]. Each possible subtree or branch of the fully grown tree is considered, starting from the bottom and iteratively moving upwards. First, however, the misclassification rate of the fully grown tree is assessed in the context of the test set, which will normally not be very good since the tree is overgrown. Then each of these subtrees is replaced by a single terminal node (which is assigned the appropriate majority class) and the misclassification rate of the entire tree is assessed again. If the misclassification rate is found to have reduced, the reduced (or pruned) tree is kept and the process is repeated until no further reduction of the misclassification rate is produced by pruning [192, p. 439].

3.9 Support vector machines

As with the previous classification methods, *support vector machines* (SVMs) may be used to predict a binary outcome of observations. The objective of SVMs is to use a single linear surface, known as a *hyperplane*, to separate the observations in the training data, belonging to the two different classes, with the largest margin of separation possible. The observations located closest to the hyperplane in the two classes are known as *support vectors*. The outcome of new observations may be predicted based on which side of the hyperplane they lie [38, p. 643], [45], [138, p. 119].

Cortes and Vapnik [45] note that in reality, however, it is very often the case that such a linear hyperplane cannot be found, in which case a further step is taken by introducing slack variables, and/or by converting the data into a higher dimensional space known as a *feature space*, with the aid of *kernels*. Within this feature space a linear hyperplane may then be placed appropriately.

3.9.1 Linear separability

In order to illustrate graphically the linearly separable case of observations in two dimensions, consider Figure 3.21. In the figure, the observations with outcome of class 1 are denoted by circles and the observations with outcome of class 2 are denoted by squares. Using SVMs the aim is to draw a single line, denoted by H, so as to separate the two types of observations with the largest possible margin of separation, denoted by M. Those observations that lie on the dashed boundary lines of the margin of separation, denoted by B_1 and B_2, are the support vectors.

Notice that other lines may also be drawn to separate the observations into two classes, as shown in Figure 3.22, but such lines achieve smaller margins of separation. In the example provided, M_2 is the larger of the two margins of separation and thus its corresponding hyperplane is favoured.

The general SVM working discussed above is now described mathematically for the linearly separable case. According to Çaydaş and Ekici [38, p. 643] and Manning *et al.* [134, p. 322], a separating hyperplane or decision hyperplane H may be defined by a vector $\boldsymbol{w}$ that is perpendicular to the hyperplane, and an intercept b (the perpendicular distance from the origin to the

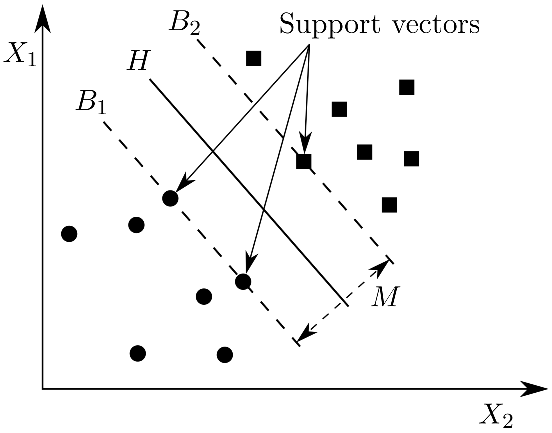


FIGURE 3.21: *SVM fundamentals.*

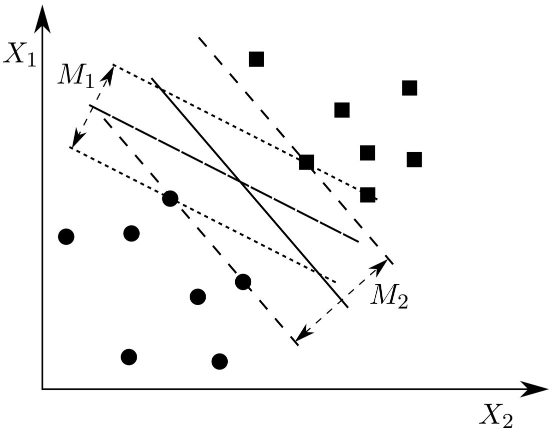


FIGURE 3.22: *SVM margin of separation.*

hyperplane). The purpose of b is to indicate which of the hyperplanes, that are perpendicular to $\boldsymbol{w}$, is selected. Since it is known that $\boldsymbol{w}$ is perpendicular to H, it follows

$$\boldsymbol{w} \cdot \boldsymbol{x}_H + b = 0 \tag{3.42}$$

for any vector $\boldsymbol{x}_H$ from the origin to a point on H. In the same fashion, points lying on the boundary lines B_1 and B_2 parallel to H, denoted by $\boldsymbol{x}_{B_1}$ and $\boldsymbol{x}_{B_2}$ respectively, will satisfy

$$\boldsymbol{w} \cdot \boldsymbol{x}_{B_1} + b = -1$$

and

$$\boldsymbol{w} \cdot \boldsymbol{x}_{B_2} + b = +1,$$

as shown in Figure 3.23.

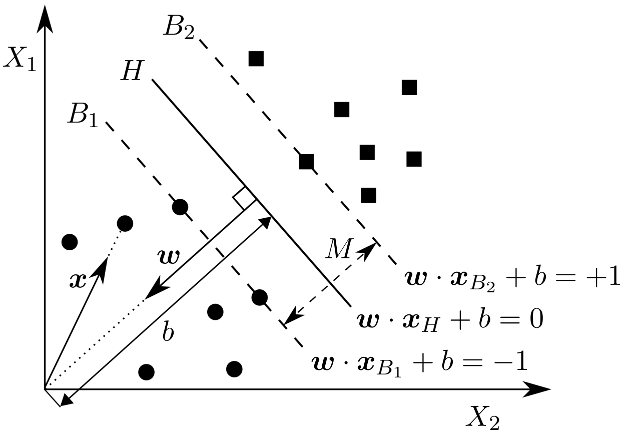

X_1
B_2
H
B_1
M
$\boldsymbol{w}$
$\boldsymbol{x}$
b
$\boldsymbol{w} \cdot \boldsymbol{x}_{B_2} + b = +1$
$\boldsymbol{w} \cdot \boldsymbol{x}_H + b = 0$
$\boldsymbol{w} \cdot \boldsymbol{x}_{B_1} + b = -1$
X_2

FIGURE 3.23: *SVM support vectors.*

Denote each of the n observations or alternatives belonging to a sample $\mathcal{N}$ by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. The observation $\boldsymbol{x}_i$ is assigned a label $y_i = +1$ for the one outcome class and $y_i = -1$ for the other. Note that although the outcome classes are binary, it is customary to use the labels +1 and −1 as opposed to 0 and 1. Based on (3.42) and using the labels +1 and −1, Welling [244, p. 2] derives two decision rules or constraints

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq +1, \quad \text{for } y_i = +1 \tag{3.43}$$

and

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1, \quad \text{for } y_i = -1. \tag{3.44}$$

By including the labels of the outcome classes into (3.43) and (3.44) they may be combined into the single constraint

$$y_i((\boldsymbol{w} \cdot \boldsymbol{x}_i) + b) - 1 \geq 0, \quad i \in \mathcal{N}. \tag{3.45}$$

Therefore,

$$y_i((\boldsymbol{w} \cdot \boldsymbol{x}_i) + b) - 1 = 0, \quad i \in \mathcal{N} \tag{3.46}$$

for all support vectors. The width of the margin of separation, m, may be determined as follows. Let $\boldsymbol{x}_i$ denote an observation on B_1 and let $\boldsymbol{x}_j$ denote an observation on B_2. The difference between these observations is denoted by $\boldsymbol{R} = \boldsymbol{x}_j - \boldsymbol{x}_i$, as shown in Figure 3.24.

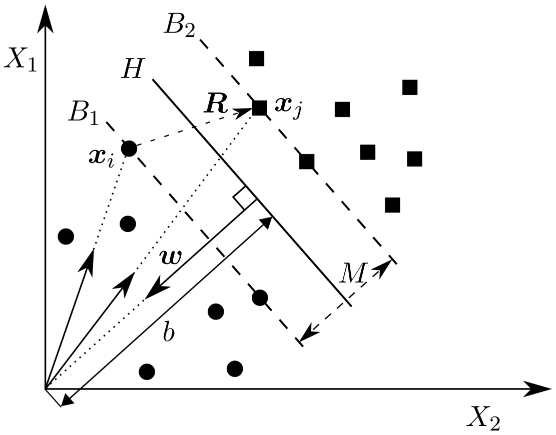


FIGURE 3.24: *Maximising the margin of separation in an SVM.*

In order to transform $\boldsymbol{R}$ so that its magnitude is m, it needs to be made perpendicular to H and thus parallel to $\boldsymbol{w}$, which may be achieved by multiplying $\boldsymbol{R}$ by the *unit vector* in the direction of $\boldsymbol{w}$. The unit vector in the direction of $\boldsymbol{w}$ may be found by diving $\boldsymbol{w}$ by its magnitude, so that

$$M = (\boldsymbol{x}_j \cdot \boldsymbol{x}_i) \cdot \frac{\boldsymbol{w}}{||\boldsymbol{w}||} \tag{3.47}$$

or

$$M = \frac{(\boldsymbol{x}_j \cdot \boldsymbol{w})}{||\boldsymbol{w}||} - \frac{(\boldsymbol{x}_i \cdot \boldsymbol{w})}{||\boldsymbol{w}||}. \tag{3.48}$$

For an observation lying on B_2, such as $\boldsymbol{x}_j$, its outcome label y_j is $+1$. Similarly for an observation lying on B_1, such as $\boldsymbol{x}_i$, $y_i = -1$. It is therefore possible to rewrite (3.46) as

$$+\big((\boldsymbol{w} \cdot \boldsymbol{x}_1) + b\big) - 1 = 0, \tag{3.49}$$

and

$$-\big((\boldsymbol{w} \cdot \boldsymbol{x}_2) + b\big) - 1 = 0. \tag{3.50}$$

By solving for $\boldsymbol{w} \cdot \boldsymbol{x}_i$, (3.49) and (3.50) become

$$\boldsymbol{w} \cdot \boldsymbol{x}_1 = 1 - b, \tag{3.51}$$

and

$$\boldsymbol{w} \cdot \boldsymbol{x}_2 = -1 - b, \tag{3.52}$$

respectively.

By substituting (3.51) and (3.52) into (3.48) it follows that

$$M = \frac{(1-b)}{||\boldsymbol{w}||} - \frac{(-1-b)}{||\boldsymbol{w}||}, \tag{3.53}$$

which may be simplified to

$$M = \frac{2}{||\boldsymbol{w}||}.$$

In order to maximise m, the quadratic programming problem

$$\begin{aligned} \min \quad & \frac{1}{2}||\boldsymbol{w}||^2, && (3.54)\\ \text{subject to} \quad & y_i\big((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\big)-1 \geq 0, \quad i \in \mathcal{N} && (3.55) \end{aligned}$$

may therefore be solved.

It is possible to go one step further and convert (3.54) and (3.55) to a *dual optimisation problem*, which is also a quadratic programming problem. This may be achieved by producing a Lagrangian

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i\big(y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)-1\big) \tag{3.56}$$

from the objective function and then maximising it with respect to the *dual variables* or the *Lagrange multipliers* $\alpha_1, \alpha_2, \ldots, \alpha_n$ [253, p. 2]. By taking the partial derivatives of (3.56) with respect to b and the components of $\boldsymbol{w}$, and setting each to zero (*i.e.* setting the gradients of L in the direction of $\boldsymbol{w}$ and b equal to zero), the equations

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \tag{3.57}$$

and

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{3.58}$$

are obtained [134, p. 324]. By substituting (3.57) into (3.56), it follows that

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i\cdot\boldsymbol{x}_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i\cdot\boldsymbol{x}_j) - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i, \tag{3.59}$$

which may further be simplified by making use of (3.58) so as to obtain

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i\cdot\boldsymbol{x}_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i\cdot\boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i. \tag{3.60}$$

After further simplification of (3.60) the dual problem

$$\begin{aligned} \max \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j), && (3.61)\\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \quad i \in \mathcal{N}, && (3.62)\\ & \alpha_i \geq 0, \quad i \in \mathcal{N}. && (3.63) \end{aligned}$$

is obtained [217, p. 447], [244, p. 2], [253, p. 2].

Again consider an arbitrary point $\boldsymbol{x}_j$ lying on B_2 and an arbitrary point $\boldsymbol{x}_i$ lying on B_1. According to Çaydaş and Ekici [38, p. 643], it is possible to determine the optimal value of b in terms of the optimal Lagrange multipliers as

$$b = -\frac{1}{2}\sum_{\text{SVs}} y_i\alpha_i\big[(\boldsymbol{x}_1\cdot\boldsymbol{x}_i)+(\boldsymbol{x}_2\cdot\boldsymbol{x}_i)\big], \tag{3.64}$$

where only the support vectors (SVs) are considered since their α-values are zero. The final decision function is then given by

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b, \tag{3.65}$$

from which an unknown data observation $\boldsymbol{x}$ may be classified, according to Figure 3.24, as

$$\boldsymbol{x}\text{ is }\begin{cases}\text{a square observation,} & \text{if } f(\boldsymbol{x})\geq 0,\\ \text{a circle observation,} & \text{otherwise.}\end{cases} \tag{3.66}$$

3.9.2 Linear non-separability

The above methodology is well suited to the case where a single linear hyperplane separates observations into two different outcome classes. In reality, however, this is rarely the case and often the type of situation depicted in Figure 3.25 occurs, where no such hyperplane may be found.

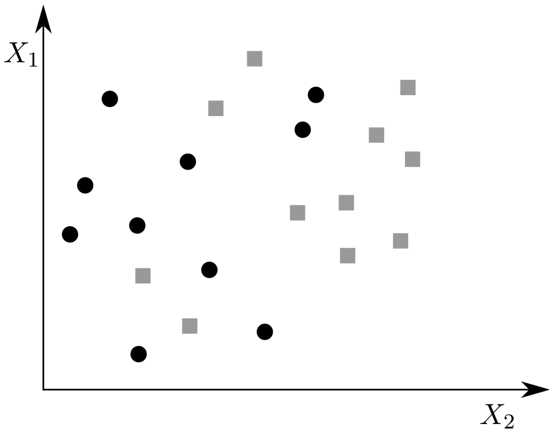


FIGURE 3.25: *No single linear hyperplane separates the observations into two outcome classes.*

In this case it is possible to relax the constraints of (3.45) by introducing a *slack variable*, denoted by $\xi_i \geq 0$, for each constraint, so that

$$y_i\big((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\big)\geq 1-\xi_i, \quad i\in\mathcal{N}. \tag{3.67}$$

Note that (3.67) may, however, be satisfied by $\boldsymbol{x}_i$ by even if it is located far on the 'wrong' side of the hyperplane. According to Manning *et al.* [134, p. 327], it is thus necessary to restrict the extent to which a point may be on the 'wrong' side by penalising the sum of ξ_i-values by an error penalty or weight parameter, denoted here by C, such that the original primal problem (3.54)–(3.55) becomes

$$\min \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n}\xi_i, \tag{3.68}$$

$$\text{subject to} \quad y_i\big((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\big)\geq 1-\xi_i, \quad i\in\mathcal{N}, \tag{3.69}$$

$$\xi_i\geq 0, \qquad i\in\mathcal{N}. \tag{3.70}$$

As before, Lagrange multipliers may be used to convert the primal problem (3.68)–(3.70) to a dual problem of the form

$$\max \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j), \tag{3.71}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad i \in \mathcal{N}, \tag{3.72}$$

$$0 \leq \alpha_i \leq C, \quad i \in \mathcal{N}, \tag{3.73}$$

from which the coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ may be found, and the decision function is again as specified in (3.65). Notice that the slack variables $\xi_1, \xi_2, \ldots, \xi_n$ are no longer present in (3.71)–(3.73), but that the value C still acts as a *regularisation* term which allows for the control of over-fitting by setting an upper bound on the values $\alpha_1, \alpha_2, \ldots, \alpha_n$.

According to Zhu [253, p. 4], it may be shown that $0 < \alpha < C$ and $\xi = 0$ for a support vector $\boldsymbol{x}_i$. Points corresponding to a value $\alpha_i = 0$ (*i.e.* points that are not support vectors) may be safely ignored without affecting the solution, and Çaydaş and Ekici [38, p. 643] note that this is partly why SVMs are largely computationally efficient: the number of support vectors is typically smaller than the total number of training data points. If, however $\alpha_i = C$, then $\boldsymbol{x}_i$ lies between the separation boundary (on its 'correct' side) and H if $\xi_i \leq 0$, or on the 'incorrect' side of H if $\xi_i > 0$. It is thus possible to think of ξ_i as the distance at which $\boldsymbol{x}_i$ lies on the wrong side of its 'correct' separation boundary. By decreasing C, the misclassification rate is therefore considered less important and subsequently more of the training data are misclassified [140].

Since the values of $\xi_1, \xi_2, \ldots, \xi_n$ and C assist in avoiding overfitting, these values may be used not only for the linearly non-separable case, but also when the data are linearly separable, so as to better partition the bulk of the data by turning a blind eye to a few of points that are misclassified [134, p. 327].

3.9.3 The use of kernels

As mentioned above, (3.68)–(3.70) may be considered the primal problem, whereas (3.71)–(3.73) is the dual problem. Although both these problems are quadratic programming problems, the primal problem has $d+1$ variables, where d denotes the number of dimensions of $\boldsymbol{x}$, whereas the dual problem has n variables. Traditionally the smaller of the two problems is selected to solve, but the generally larger one — in this case the dual problem — allows for the use of kernels [138, p. 126], [253, p. 2].

Kernels may be employed in conjunction with any SVM (in addition to incorporating slack variables), even for those cases where a single linear hyperplane separating observations of two different outcome classes cannot be found. The fact that the dual problem of (3.71)–(3.73) and the decision function of (3.65) rely on the inner product $\boldsymbol{x}_i \cdot \boldsymbol{x}_j$ of two points is vital as the use of kernels involves transforming the data into a higher dimensional feature space by defining a kernel function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ that produces an inner product. It is then possible to replace each $\boldsymbol{x}_i$ in (3.71)–(3.73) and (3.65) with a non-linear decision function, denoted by $\phi(\boldsymbol{x}_i)$, such that $\phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, according to which the dual problem may be solved. The result is equivalent to using a non-linear hyperplane H in the original space [38, p. 644].

Hill and Lewicki [92] point out that the class of radial basis function kernels is by far the most popular employed in conjunction with SVMs, of which the best known type is known as the

Gaussian radial basis function kernel. According to Çaydaş and Ekici [38, p. 643] and Manning *et al.* [134, p. 333], the Gaussian radial basic function kernel may be defined as

$$\phi(\boldsymbol{x}_i) \cdot \phi(\boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{2\sigma^2}\right),$$

where $\phi(\boldsymbol{x})$ is the non-linear decision function mentioned above [38, p. 644]. Note that the larger σ, the further the 'reach' and influence of the training observations. In particular, when $\boldsymbol{x}_j$ is a support vector, and σ is large, then the class of $\boldsymbol{x}_j$ will be very influential in deciding the class of $\boldsymbol{x}_i$, even if they are far apart. The converse is also true [173].

The Gaussian radial basis function, as well as some other well-known inner product kernels, is presented in Table 3.5 [89], [101, p. 6], [138, p. 127], [195, p. 140], [208].

Table 3.5: *Inner product kernel function examples.*

Type of kernel	Function
Gaussian radial basis function	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lvert\lvert\boldsymbol{x}_i - \boldsymbol{x}_j\rvert\rvert^2}{2\sigma^2}\right)$
Exponential radial basis function	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lvert\lvert\boldsymbol{x}_i - \boldsymbol{x}_j\rvert\rvert}{2\sigma^2}\right)$
Laplace radial basis function	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lvert\lvert\boldsymbol{x}_i - \boldsymbol{x}_j\rvert\rvert^2}{\sigma}\right)$
Polynomial	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\delta(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + \theta)^d$
Hyperbolic tangent (Sigmoid)	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\delta(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + \theta)$
Bessel function	$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{J_{v+1}(\sigma\lvert\lvert\boldsymbol{x}_i - \boldsymbol{x}_j\rvert\rvert)}{\lvert\lvert\boldsymbol{x}_i - \boldsymbol{x}_j\rvert\rvert^{-n(v+1)}}$

In the event that no prior knowledge about the data is available, the general-purpose Gaussian, exponential, and Laplace radial basis function kernels and the Bessel kernel are well suited. The popular polynomial kernel is well suited for data that have been normalised [101, p. 6], [140].

3.10 Statistical and machine learning assumptions

Various statistical and machine learning assumptions are considered in this section. First, the assumptions associated with regression, in general, and linear regression, in particular, are elaborated upon. This is followed by a discussion of the specific assumptions underlying the five other statistical and machine learning models considered in §3.4–§3.9, namely logistic regression, classification and regression trees, random forests, support vector machines, and the C4.5 algorithm.

3.10.1 Assumptions associated with regression in general

In order to apply the technique of linear or multiple regression to a given set of data, a number of assumptions have to be made. Berry [19] states that if an analyst has a true understanding of the assumptions of regression it will enable him or her to adequately appreciate the strengths and weaknesses of estimates and thus make the required model improvements where appropriate.

Osborn and Waters [166] conducted research in which they highlighted the three assumptions of multiple regression which should always be borne in mind. The article has since received more than 477 000 views and is well cited in the literature. The three assumptions involve the notions of normality, measurement error, and homoscedasticity. In reality, these assumptions are rarely met, but it should nevertheless be considered whether these assumptions are reasonable in the context in which the regression analysis is applied.

Normality

The first of the assumptions identified by Osborn and Waters [166] is that of normality. The variables employed should be normally distributed, *i.e.* may not be skewed[2] or kurtosis variables[3], as this may misrepresent relationships and the results of any significance tests.

Non-normality can easily be identified visually. Little and Silal [125, p. 24] suggest that the assumption may be validated by constructing a histogram plot of the residuals, which should display the classic bell-shaped curve associated with the normal distribution. Another possibility is to construct a Normal Q-Q plot[4].

If the assumption of normality is not met, there are a number of methods which may be employed to rectify this deficiency. If non-normality is present in a small sample, for example, the problem may be rectified by increasing the sample size. This is, however, not always possible, in which case Little and Silal [125, p. 24] suggest transforming the dependent variables (*e.g.* by considering the variables $\log Y$ or $\sqrt{Y}$ instead of Y), while other times it may be sufficient to include an X^2 term in the model, thus creating a quadratic relationship, or by utilising a non-linear function which describes the variables' relationships more accurately.

There are, however, authors who believe that the non-normality of residuals would not necessarily negatively influence any statistical interference [143, p. 83].

Measurement errors

Recall that the problem of measurement errors was discussed in §3.2.1. This discussion included a description of methods that may be implemented to reduce the possibility of measurement error. In standard regression models, such as linear regression, it is assumed that the values of the independent variables were collected, measured or observed without error (*i.e.* reliably). If this is not the case, or if there is correlation present between the different measurement error groups for each of the independent variables, the possibility exists that the coefficient estimates in (3.74) and (3.75) may be biased to either side.

Homoscedasticity of residuals

The presence of *homoscedasticity*, also referred to as *homogeneity of variances*, implies that the residuals exhibit an approximately constant variance across all levels of the independent variables. If this is not the case, the undesirable effect of *heteroscedasticity* will be present [6, p. 25]. Recall that the concept of residuals (as opposed to true error terms) was discussed in §3.1.3.

One way of testing for homoscedasticity, according to Osborn and Waters [166, p. 4], involves constructing a scatter plot of the standardised residuals against the predicted dependent variables of observations. For homoscedasticity to be present, the points in such a plot should be spread randomly around the zero horizontal line. Figure 3.26(a) contains an example of such a plot in the case of homoscedasticity, while Figures 3.26(b)–(c) contain examples of the presence of heteroscedasticity [246, p. 1308].

[2] A measure of the lack of symmetry of the distribution.

[3] A measure of how heavy-tailed the distribution is relative to a normal distribution.

[4] A plot of the quantiles of a data set against a given distribution.

3.10.2 Assumptions associated with linear and multiple regression

Besides the three general assumptions associated with regression identified by Osborn and Waters [166], and described above, five further assumptions which are applicable specifically to linear regression have been identified by other authors. These assumptions involve the notions of linearity, independence of residuals, multicollinearity, outliers and minimum sample size.

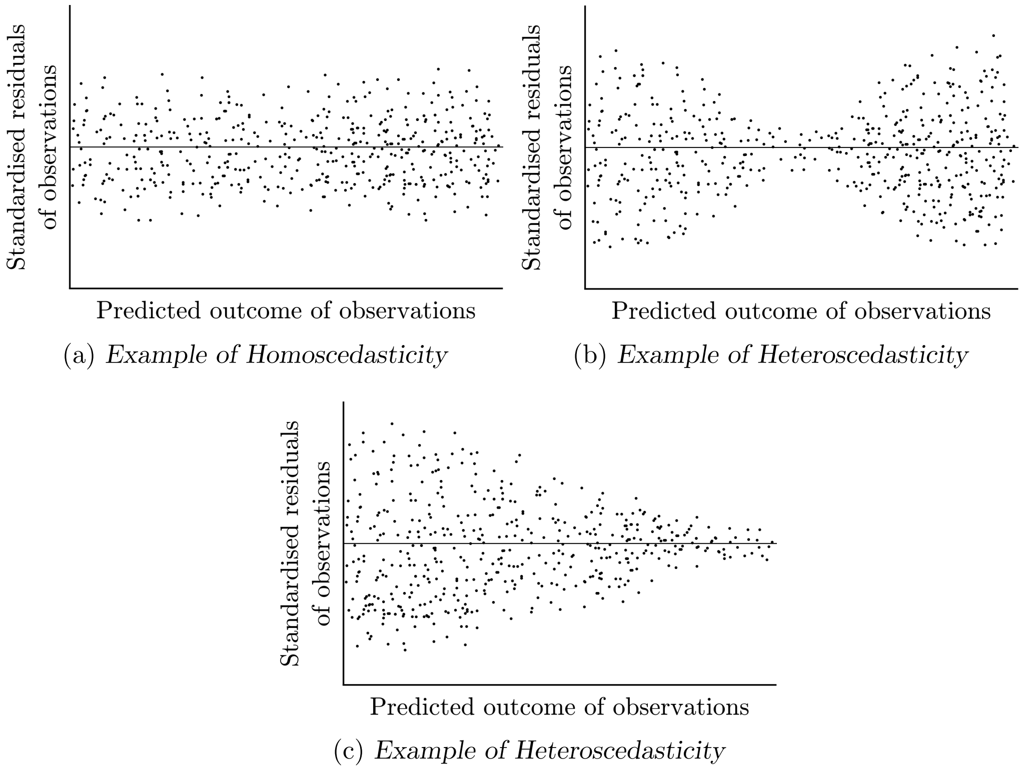
Linearity

Consider again the form of a simple linear regression equation,

$$Y = \beta_0 + \beta_1 X, \tag{3.74}$$

and of a multiple linear regression equation,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m. \tag{3.75}$$



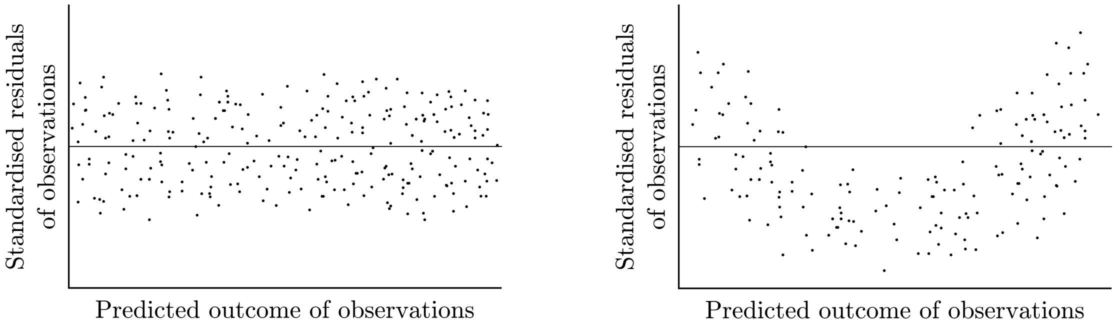
(a) *Example of Homoscedasticity*

(b) *Example of Heteroscedasticity*

(c) *Example of Heteroscedasticity*

Figure 3.26: *Plots of predicted dependent variables versus standardised residuals for three separate example sets so as to illustrate the notions of homoscedasticity and heteroscedasticity.*

The first of the five assumptions mentioned above, linearity, was also identified by Osborn and Waters [166, p. 2], who state that it will only be possible to accurately estimate the coefficients $\beta_0, \ldots, \beta_m$ in (3.74) and (3.75) if the relationship between the dependent and independent variables are truly linear in nature. It is possible to identify non-linearity by plotting the standardised residuals against the predicted dependent variables of observations. If the relationship between the dependent and independent variables is linear, the plot will take the form illustrated graphically in Figure 3.27(a), while if the relationship is curvilinear, the plot will take the form illustrated graphically in Figure 3.27(b).



(a) *Example of a linear relationship* (b) *Example of a curvilinear relationship*

FIGURE 3.27: *Plots of predicted dependent variables versus standardised residuals for two separate example sets so as to illustrate the notions of linear and curvilinear relationships.*

Independence of residuals

The assumption of *independence of residuals*, also sometimes referred to as *statistical independence* or merely *independence*, requires that the residuals — not the raw observations — are independent of each other. If this is not the case, the residuals are said to be *autocorrelated*, also known as exhibiting *lagged correlation* or *serial correlation* [6, p. 25].

According to Ayyangar [14, p. 3], one way of testing for autocorrelation involves computation of the *Durbin-Watson coefficient* (DW). The DW statistic falls in the range of $[0, 4]$, where value 2 represents no autocorrelation, and a value between 1.5 and 2.5 is deemed acceptable.

Let R_i represent the residual (difference between the observed and predicted outcome) of observation i in a sample of n observations. The Durbin-Watson test statistic is then given by

$$\mathrm{DW} = \frac{\sum_{i=2}^{n}(R_i - R_{i-1})^2}{\sum_{i=1}^{n} R_i^2}. \tag{3.76}$$

Risk of the assumption of independence of residuals being violated is mostly of concern in the context of time series data or longitudinal data, such as stock market prices (where a stock's current price is related to that of the previous and that of the following day).

Multicollinearity

Little and Silal [125, p. 33] also identified two other assumptions of linear regression which have to be considered. The first of these involves the notion of collinearity among independent variables.

In multiple regression two or more independent variables are present. According to Winston [246, p. 1322], *collinearity* exists between two of the independent variables within a multiple regression model if a positive or negative linear relationship exists between them. If more than two of the independent variables within a multiple regression model exhibit collinearity with each other, the phenomenon is referred to as *multicollinearity* [99, p. 101].

Graham [80, p. 2809] states that multicollinearity within a model places the credibility of the analysis, and the interpretation of the model and its individual $\hat{\boldsymbol{\beta}}$ estimates, at risk. The statistical significance results, such as the R^2 and p-values, of the overall model, as well as its fit and predictive ability may, however, not be negatively affected.

Montgomery and Runger [154, p. 489] claim that the effect of multicollinearity may not be too negative and that reasonable estimations may still be obtained, provided that predictions are performed using independent variable values from the original region in which the multicollinearity is present, referred to as *interpolations*. If, however, these predictions result from extrapolations (estimated using independent variable values outside the original region), the results obtained may be poor.

A specific independent variable may be highly correlated with a second independent variable, in which case it may seem as if the first variable does not significantly assist the model in predicting the dependent variable. This may be due to the fact that the first variable is predicted to a large extent by the second variable. In such a case it may be that both independent variables are good predictors in their own right and it may therefore be deemed unnecessary to include both in the model [2, p. 138].

Multicollinearity is rarely not present in a multiple regression model. Approaches toward dealing with the notion of multicollinearity are therefore not merely aimed at determining whether multicollinearity is present, but to estimate the degree to which it is present. A well-known approach toward identifying multicollinearity among independent variables involves determining the *variance inflation factor* (VIF) for each independent variable.

The use of VIFs makes it possible to quantify the negative impact of multicollinearity. O'Brien [162, p. 683] claims that the VIF is a well-known indicator of the extent to which each independent variable promotes multicollinearity in a model. According to Montgomery and Runger [154, p. 490], it is possible to quantify the extent to which each of the n independent variables contributes to multicollinearity within a regression model by determining the VIF for the estimated coefficients of each independent variable as

$$\mathrm{VIF}(\hat{\boldsymbol{\beta}}_i) = \frac{1}{(1 - R_i^2)}, \quad \text{for } i = 1, 2, \ldots, n. \tag{3.77}$$

Each R_i^2-value in (3.77) is obtained by performing regression where the i-th independent variable is used as the dependent variable and the remaining $i - 1$ independent variables remain the independent variables. In other words each individual independent variable is removed from the model in turn and taken as the dependent variable in a statistical test. According to Tu *et al.* [232, p. 458], the use of auxiliary regression in conjunction with the VIF, as described above, is the best known method for detecting multicollinearity.

A VIF value of, say 8, for the i-th independent variable may be interpreted as meaning that because the i-th independent variable is not truly independent of the other independent variables, the variance of the i-th regression coefficient is eight times larger than it would have been had the i-th variable been independent [162, p. 684].

A VIF value of 1 indicates that no multicollinearity is present, while a value of 10 is typically taken as a rule-of-thumb indicating that the specific independent variable is associated with severe multicollinearity within the model [162, p. 673]. Some authors, however, believe that a smaller VIF rule-of-thumb value of 4 or 5 should be employed [72, p. 45].

Although no easy solution exists for the problem of multicollinearity, three methods may be employed to rectify the negative effects of multicollinearity. The first method is due to Montgomery and Runger [154, p. 490], who propose that the variable believed to be causing the multicollinearity should be deleted. This option, however, comes at the cost of losing all the valuable data associated with the variable being deleted.

A second option is applicable to the case where two independent variables are highly correlated and both are good predictors, but it does not make sense to use both variables and deleting one of them is not desirable. These two variables may then be combined into a single variable [162, p. 684].

Finally, a third option may be to augment the data with new observations, if possible, in the hope of removing the dependencies which currently exist [154, p. 490].

Outliers and influential observations

Another general concern related to simple linear and multiple linear regression models raised by Little and Silal [125, p. 33] is that outliers and influential observations should have been identified and dealt with accordingly before a regression analysis is performed on the data. The definition of, as well as the identification and possible methods for dealing with, outliers and influential observations has already been discussed in §2.2.4.

Minimum sample size

Although many other suggestions exist in the literature with respect to the minimum sample size required for simple linear or multiple linear regression, the number supported by many authors is that a sample of between 25 and 30 observations per independent variable should yield satisfactory results [93]. There are some authors, however, who favour the use of larger minimum samples sizes. Methods suggested for determining the minimum sample size include the use of predetermined tables (Knofczynski and Mundfrom [110] and Murphy and Myors [159]) or specially designed software (Faul *et al.* [65]).

3.10.3 Assumptions associated with logistic regression

Although a substantial number of assumptions have to be satisfied when applying linear or multiple linear regression, far fewer assumptions have to be satisfied when applying logistic regression [175]. According to Spicer [209, p. 135], the classical assumptions of normality (discussed in §3.10.1), homoscedasticity (also discussed in §3.10.1), and linearity (discussed in §3.10.2) need not be satisfied in the case of logistic regression. The distribution of the residuals should rather resemble those of a binomial distribution. For large samples, the binomial distribution approximates the normal distribution [143, p. 83].

In all three cases, however, if the above assumptions are indeed satisfied, the possibility of an improved model increases. According to Park [170, p. 157], there are still four assumptions that have to be satisfied in order for a logistic regression analysis to be justified. These four assumptions involve mutual exclusivity of categorical dependent variables, the method of coding dependent variables, and the notions of multicollinearity, and of linearity of the logit.

The first requirement of logistic regression is the presence of categorical dependent variables. The groups into which these dependent variables are categorised should also be mutually exclusive [10]. Consider, for example, the case in which the success of a tertiary student has to be predicted. One possibility may be to create three mutually exclusive groups containing students who (a) failed, (b) completed their degrees in longer than the minimum time, and (c) completed their degrees in minimum time. If, however, the three groups were classified as students who (a) failed, (b) completed their degrees, and (c) completed their degrees in minimum time, the students in the 'completed their degrees' and 'completed their degrees in minimum time' groups would overlap. This would imply that the two groups are not mutually exclusive.

Logistic regression allows the analyst to estimate the probability of a certain event occurring. The dependent variable should therefore be coded appropriately such that $P(Y = 1)$ may be interpreted correctly, where Y represents the dependent variable. The desired outcome should be given the label of 1 [170, p. 157].

As discussed in §3.10.2, the independent variables should not be highly correlated with one another, as was the requirement for linear regression [10].

Unlike the linearity assumption discussed in §3.10.2 (which is not applicable to logistic regression), the linearity of the logit assumption requires that the independent variables should be linearly related to the logit transform of the dependent variable [170, p. 157]. According to Hilbe [91, p. 83], it is possible to assess the linearity of the logit assumption by use of the *Box-Tidwell* method. This method involves adding all quantitative independent variables (as only quantitative variables are applicable for this assumption), as well as the natural log transform of each quantitative independent variable into a logistic regression model. For example, assume that the age of individuals are being used as the independent variable to assess whether or not those same individuals own their own vehicle. The model may then be defined as $\mathit{Is\ vehicle\ owner} = \mathit{Age} \times \ln(\mathit{Age})$. If any of the log transform variables, in this case $\ln(\mathit{Age})$, are shown to be statistically significant, it may be assumed that the assumption of linearity of the logit has been violated. The violation of this assumption may result in biased parameter estimations and biased standard errors.

In addition to these four assumptions another four assumptions, applicable to logistic regression, have been identified by other authors [72, p. 46], [168, p. 161], [236]. These four assumptions involve the presence of outliers, independence of residuals, as well as concerns in terms of limited measurement error and minimum sample size.

Pallant [168, p. 161] notes the importance of testing for outliers, as well as for influential observations, when applying logistic regression as the presence of outliers in the training set might result in incorrect prediction of new data. The definition of, and the identification and possible methods for dealing with, outliers and influential observations have already been discussed in §2.2.4.

According to Garson [72, p. 46], the independence of residuals and thus a lack of autocorrelation, as discussed in §3.10.2, is also assumed for a logistic regression model. Again, the Durbin-Watson test (3.76) may be used to assess this assumption.

The third of the four assumptions applicable to logistic regression is the assumption of limited measurement error discussed in §3.10.1.

The last of the assumptions justifying the use of logistic regression is that an adequate sample size should be used. Van der Ploeg *et al.* [236] are of the opinion that as few as 20 observations are sufficient for each independent variable. Moisen [152, p. 585], on the other hand, believes that for classification problems in which two outcomes are possible, a minimum of 200 observations are required.

Another easy guideline to follow, suggested by Peduzzi *et al.* [174], is that the minimum number of observations in a sample for multiple logistic regression may be calculated as

$$\text{Minimum number} = \frac{10m}{p}, \tag{3.78}$$

where m is the number of independent variables and p is the proportion of observations whose outcome is classified into the smallest class.

3.10.4 Assumptions associated with classification and regression trees

A major advantage of using CART is that almost all of the traditional assumptions required by the parametric methods mentioned above do not apply. This is due to CART being a non-parametric modelling approach, thereby rendering redundant any assumptions about the distributions of either the dependent or independent variables. In addition, the four assumptions of homoscedasticity (§3.10.1), multicollinearity (§3.10.2), the distribution of residuals, and the accommodation of outliers (§3.10.2) do not apply to CART [160, p. 278], [226, p. 22], [252, p. 10].

The assumptions involving measurement error (§3.10.1), mutually exclusive dependent variable categories, and minimum sample size do, however, apply. Li *et al.* [120, p. 977] suggest a minimum sample of 240 samples for the smallest class so as to obtain stable results.

3.10.5 Assumptions associated with random forests

The method of random forests entails fitting many classification trees and as a result its assumptions are the same as those of CART mentioned above [48, p. 2784], [96, p. 2682].

Li *et al.* [120, p. 977] mention that typically classification tree algorithms are most affected by sample size, but not random forests. Cutler *et al.* [48, p. 2788–2790] conducted an analysis using what they admit was a 'smaller' sample size of 107 observations, but were satisfied with both the predictive accuracy of 83.1% and the expected variable importance scores. Thus, although a larger sample size might have produced better results, it was shown that it is possible to produce satisfactory results and expected variable importance scores with a sample size of 107.

3.10.6 Assumptions associated with support vector machines

The modelling approach of SVMs, like CART, is non-parametric and hence no assumptions are made about the distributions of either the dependent or independent variables, nor are any of the other traditional assumptions required.

The method of SVMs requires the same assumptions as CART (§3.10.4), with one exception. Unlike random forests and CART, which are robust in respect of outliers, Debruyne [52, p. 2] states that the method of SVMs is sensitive to outliers, as well as influential observations. A methodology for handling outliers and influential observations has to be considered (see §3.10.2).

In addition, the assumption of minimum sample size is applicable. In 2014, Li *et al.* [120, p. 977] conducted research on the effect of sample size on a number of machine learning algorithms. They found that SVMs are least affected by small sample sizes and that a sample size of the smallest class greater than 60 will produce stable results in the sense that increasing the sample size from that point does not significantly alter the results produced.

3.10.7 Assumptions associated with the C4.5 algorithm

According to Stärk and Pfeiffer [210, p. 1], the decision tree algorithm known as C4.5 is also a non-parametric approach and hence the same assumptions apply as for CART (see §3.10.4). In addition, the algorithm has a built-in step which removes suspected outliers [184]. The outlier assumption is therefore not applicable.

The algorithm does, however, require that a minimum sample size cut-off be identified. Li *et al.* [120, p. 977] found the C4.5 algorithm to be very sensitive to a small sample size and suggested that a sample size of the smallest class of 240 plus is required to obtain stable results.

3.11 Chapter summary

Various concepts of basic statistics, statistical learning, statistical regression, and statistical classification were reviewed in this chapter.

In §3.1, a number of general statistical and statistical learning concepts were presented. It started out by discussing the concepts of inputs and outputs of a process, and the relationship which exists between them. Thereafter, simple linear regression and multiple linear regression were presented as examples of the relationship between these inputs and outputs. The modelling paradigms of explanatory modelling and predictive modelling were discussed thereafter. This was followed by a presentation of the difference between the two main methods of estimating outputs based on inputs, namely parametric methods and non-parametric methods. Next, the themes of the predictive accuracy of a model *versus* its interpretability, supervised statistical learning *versus* unsupervised statistical learning, and regression problems *versus* classification problems were considered. The distinction between experimental research and non-experimental research, as well as the difference between internal and external research, were discussed thereafter. The final part of the section was dedicated to presenting, with the aid of examples, the two data splitting methods of cross-validation and bootstrapping.

In §3.2, the four possible reasons for a correlational relationship between input and output variables were presented, namely: bias caused by measurement or selection error, chance relationships, confounders, and causal relationships.

It was explained in §3.3 why linear regression is not appropriate for classification, after which the fundamental principles of five alternative models which are indeed appropriate for this purpose were discussed in turn, namely logistic regression (§3.4), CART (§3.6), random forests (§3.7), the C4.5 algorithm (§3.8), and SVMs (§3.9).

In §3.10, a discussion followed on the assumptions associated with regression in general, linear and multiple regression, and the five models mentioned above. By combining the recommendations of the various sources cited in §3.10.1–§3.10.7 with respect to assumptions underlying specific methods of analysis, a summary of the assumptions may be found in Table 3.6. Ideally all the assumptions of a model should be satisfied so as to justify its use. As explained in §3.1.4, however, the particular aim of a study should be taken into consideration. If the aim is prediction, and not the search for statistically significant variables, then for the last five models listed in Table 3.6, the only assumptions required to justify their use are the eight assumptions of independence of residuals, measurement errors, outliers, influential observations, mutually exclusive categorical dependent variables, minimum sample size, correct coding of dependent variables, and linearity of the logit.

Table 3.6: *Assumptions associated with various methods of statistical analysis.*

Assumption	Linear regression	Multiple linear regression	Logistic regression	CART	Random forests	SVMs	C4.5
Linearity	✓	✓					
Independence of residuals	✓	✓	✓				
Homoscedasticity of residuals	✓	✓					
Normally distributed residuals	✓	✓					
Measurement errors	✓	✓	✓	✓	✓	✓	✓
Limited multicollinearity		✓	✓				
Handling of outlier observations	✓	✓	✓			✓	
Handling of influential observations	✓	✓	✓			✓	
Mutually exclusive categorical dependent variables			✓	✓	✓	✓	✓
Minimum sample size	✓	✓	✓	✓	✓	✓	✓
Correct coding of dependent variable			✓				
Linearity of the logit			✓				

CHAPTER 4

Ensemble learning and multi-criteria decision analysis

Contents

4.1 Ensemble learning 89
4.2 Integration strategies in ensemble learning 91
4.3 Multi-criteria decision making 93
4.4 Chapter summary 106

In this chapter, the literature on two subdisciplines within the intersection between statistics and operations research, known as *ensemble learning* and *multi-criteria decision analysis* (MCDA), are reviewed. First, a background on ensemble learning is presented, and this is followed by a description of the three types of mechanisms that may be utilised when constructing an ensemble method. The mechanism of the third type of ensemble construction is then discussed in more detail, as it is applied later in this thesis. The final part of the ensemble learning section entails a discussion on the two available types of integration strategies used for combining base model predictions, namely *static integration* and *dynamic integration*.

The second part of the chapter is dedicated to MCDA. The different manners in which MCDA methods may be classified are discussed, and this is followed by a description of how a typical MCDA problem may be formulated. Next, twenty eight well-known MCDA methods are presented briefly, of which ELECTRE III is elaborated upon in more detail due to its importance within this thesis. Finally, a framework is presented for selecting MCDA methods based on the specific decision situation to which they are applied. The chapter closes with a brief summary of the contents of the chapter.

4.1 Ensemble learning

The process of combining the statistical predictions obtained from multiple models in a pre-specified manner, so as to obtain an overall prediction, is known as *ensemble learning*. According to Dietterich [54, p. 1], Merz [146, p. 193], and Opitz and Maclin [165, p. 169], it has been demonstrated that the predictive ability of ensemble models is often superior to the predictive ability of any single constituent classifier model. Ensemble models function based on the notion of exploiting the complementary strength of different learning models [251, p. 5].

4.1.1 Mechanisms for constructing ensemble methods

According to Kotsiantis [113, p. 263], there are three main strategies according to which ensemble models may be constructed:

- using a single learning model, but applying it to many different subsets of the original training data,
- using a single learning model, but applying it in conjunction with varying training parameters, or
- combining the use of different learning models.

The methods of *bagging* (see §3.6.4) and *boosting* (see §3.6.4) are examples of the first type of construction method mentioned above [88, p. 605]. An example of the second involves varying the initial weights of each of a number of neural networks that form part of an ensemble. While the first two of the above-mentioned construction methods involve using the same base learning methods, the third involves a combination of the outputs of different models so as to obtain a single, final prediction for each of a number of observations, *e.g.* combining the output from random forests and neural networks, but using the same training data [250, p. 11].

Although the vast majority of the literature on ensemble methods is focused on the first two construction methods mentioned above, the remainder of this section is dedicated to the third method which involves a combination of different learning models in the context of the same training data.

4.1.2 Ensemble construction

The mechanism of the third type of ensemble construction methods described in §4.1.1 is illustrated graphically in Figure 4.1, and is described in some detail in this section. The various models used in such a mechanism are henceforth referred to as the *base models*, while the overarching model produced by combining the base models according to a certain weighting scheme is called the *combined weighted model*.

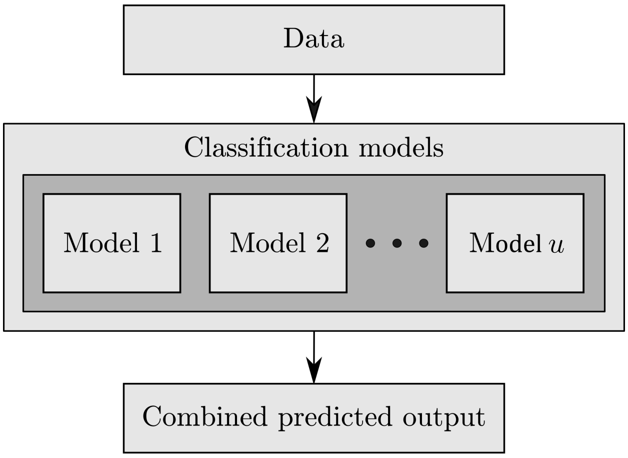


Figure 4.1: *An ensemble method which combines output from u different models.*

Examples in the literature of this type of ensemble method include the models proposed by Bhanot *et al.* [21], Hassan *et al.* [87, p. 2], Kedarisetti *et al.* [102], Yang *et al.* [249], and Yang *et al.* [251].

It is sometimes the case in the literature that, along with base models being combined, the process also involves finding a subset of model variables that maximises the prediction of the output so as to achieve an improved accuracy. The vast majority of such cases furthermore involve reducing the number of variables from thousands to a more manageable number. An example of such a method was proposed by Hassan *et al.* [87], who initially started out with 25 000 genes, but only sought a few that are highly effective in predicting a certain type of cancer. This thesis is not concerned with reducing the number of variables to a manageable size from a very large number. For that reason, the ensemble models discussed in this review are focussed on increasing the accuracy of predictions, not on diminishing the number of variables.

4.1.3 Base model selection

Yang *et al.* [251, p. 8] identified the step of selecting base models to enter an ensemble model as vital. The selected models have to satisfy certain criteria according to Hansen and Salamon [85]. More specifically, they have to be both *accurate* and *diverse*. Accuracy refers to the requirement that they should predict outcomes better than random estimations (*i.e.* better than 50% predictive accuracy) and diversity implies that the various models should make prediction errors on different observations. In addition to these two characteristics, Yang *et al.* [251, p. 8] also identified *computational efficiency* of a base model as an important selection factor.

4.1.4 Disadvantages of ensemble methods

Although ensemble models have produced results of improved accuracy in various studies, Kotsiantis [113, p. 264] and Yang *et al.* [250, p. 12] identified two disadvantages of their use. The first is that the increased accuracy is accompanied by the cost of having a more complex model with less comprehensibility. This may prove to be an obstacle for a non-expert user in terms of making sense of the reasoning of the ensemble model's results. Secondly, a complex ensemble model often requires both more computational power and memory storage capacity than a single base model.

4.2 Integration strategies in ensemble learning

Once the base models of an ensemble have been identified, the choice of which base model's prediction will be used, or how these models are to be combined, has to be considered. This is often a challenge as it is not obvious how the integration should be achieved [181, p. 2]. The manner in which this task is performed is referred to as the *integration strategy*.

4.2.1 Static integration strategies

According to Puuronen *et al.* [181, p. 3] and Woods *et al.* [247, p. 405], two main types of integration strategies exist: the first is known as *selecting* and the second as *combining* or *classifier fusion*. Selecting methods involves selection of a single base model, considered to perform best, whose predictions are then taken as the predictions of the entire ensemble model. The method of combining models, on the other hand, involves integrating the predictions of all the base models in a specified manner so as to attempt to achieve consensus among the base models. Examples of three popular methods of selecting and combining integration models are discussed in this section.

The integration strategy known as *cross-validation majority* (CVM) is an example of a selecting integration strategy. CVM, according to Merz [145], utilises cross-validation to assess the predictive ability of each model, the results of which are used to select the single best base model. The process of cross-validation was reviewed in §3.1.11.

The first and simplest combining integration strategy is known as *simple voting*, *majority voting* or *select all majority*. It involves combining the output of the different models in an unweighted averaging manner, *i.e.* the predictions of each base model are considered to have equal importance. For each observation, the class that receives the most votes from the various base models is selected [102, p. 985], [230, p. 59], [248, p. 422].

An extension of simple voting, sometimes also referred to in the literature as *majority voting*, is referred to as *weighted majority voting* or *weighted voting* in this thesis. Weighted voting is also a combining integration strategy. Instead of each base model being assigned equal importance in the final prediction of an observation, each base model is assigned a weight based on its predictive competence when predicting in the context of a test set. Thus, as before, the predictions from all the base models are combined, but now final prediction is determined as the weighted majority vote for each observation. In this manner, these models with more predictive power have more influence on the final prediction. It has been found that weighted majority voting generally produces better results than those of majority voting [21, p. 596], [230, p. 59].

4.2.2 Dynamic integration strategies

Besides being partitioned into the classes of selecting or combination methods, ensemble selection strategies may also be classified as either *static* or *dynamic* [231, p. 280]. According to Tsymbal *et al.* [230, p. 59], static integration methods consider all the observations simultaneously in one observation space, and once a decision has been reached on how to integrate the base models, it is applied to all the observations uniformly. Dynamic integration methods, on the other hand, combine the base model predictions differently for each observation. These methods are dynamic in the sense that the integration of the base models changes continually. An example of the two types of weighting may be visualised as follows. Consider the case of n alternatives and u base models (BMs). There would then exist an $u \times n$ matrix containing predictions by each base model for each alternative. In such a case, static weighting would entail determining a single prediction (Pred) for each base model, but for all the alternatives, as presented in Table 4.1. For the same scenario, an example of dynamic weighting may be seen in Table 4.2. Unlike static weighting, which determines a single weight for all the alternatives, dynamic weighting produces a weight for each base model as it corresponds to each alternative.

Table 4.1: *Example of static weighting.*

Alter.	BM$_1$	BM$_2$	$\cdots$	BM$_u$
1	Pred$_{11}$	Pred$_{12}$	$\cdots$	Pred$_{1u}$
2	Pred$_{21}$	Pred$_{22}$	$\cdots$	Pred$_{2u}$
3	Pred$_{31}$	Pred$_{32}$	$\cdots$	Pred$_{3u}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Pred$_{n1}$	Pred$_{n2}$	$\cdots$	Pred$_{nu}$
Weight	W_1	W_2	$\cdots$	W_u

Table 4.2: *Example of dynamic weighting.*

Alter.	BM$_1$	BM$_2$	$\cdots$	BM$_u$
1	Pred$_{11}$	Pred$_{12}$	$\cdots$	Pred$_{1u}$
Weight$_1$	w_{11}	w_{12}	$\cdots$	w_{1u}
2	Pred$_{21}$	Pred$_{22}$	$\cdots$	Pred$_{2u}$
Weight$_2$	w_{21}	w_{22}	$\cdots$	w_{2u}
3	Pred$_{31}$	Pred$_{32}$	$\cdots$	Pred$_{3u}$
Weight$_3$	w_{31}	w_{32}	$\cdots$	w_{3u}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Pred$_{n1}$	Pred$_{n2}$	$\cdots$	Pred$_{nu}$
Weight$_n$	w_{n1}	w_{n2}	$\cdots$	w_{nu}

4.3 Multi-criteria decision making

A well-known branch of the theory of decision making is *multi-criteria decision making* (MCDM), also known as *multi-criteria decision analysis* (MCDA). MCDA represents a family of methods which may be utilised to weigh up a number of possible alternatives or choices in the presence of various, possibly conflicting, decision criteria or objectives, searching for solutions that embody suitable confliction trade-offs [43, p. 1201].

In the context of MCDA, a decision maker may be defined as an entity who is responsible for decision making. This entity may take the form of either a single individual, a group of individuals, or an organisation. A decision maker seeks to obtain good recommendations for a particular scenario or problem, known as a *decision making situation.* According to Eldrandaly *et al.* [63], the decision maker should be able to either implicitly or explicitly provide the required information pertaining to decision preferences. It is thus possible to claim, according to Guitouni and Martel [82, p. 503], that MCDA revolves around the idea of abandoning the search for an *optimal solution* in exchange for one that is deemed *satisfactory* by the decision maker.

4.3.1 Classifying MCDA methods

MCDA methods have previously been classified in various manners in the literature. Mendoza and Martins [144, p. 2] mention some of the most popular classification methodologies.

The first classification method involves classifying MCDA methods as those which are either considered part of *multi-objective decision making* (MODM) or part of *multi-attribute decision making* (MADM). The main difference between the two paradigms is that MODM is aimed at selecting an alternative or subset of alternatives from a very large set, while MADM involves selecting an alternative or subset of alternatives from a smaller, finite set [43, p. 1201], [157, p. 26]. The distinction between MODM and MADM is made more precise by eight metacriteria proposed by Malczewski [131]. These metacriteria are listed in Table 4.3.

Table 4.3: *Metacriteria for distinguishing between MODM and MADM methods, as presented by Malczewski [131].*

Criteria for identifying	MODM	MADM
Criteria defined by	Objectives	Attributes
Constraints defined	Explicitly	Implicitly
Alternatives defined	Implicitly	Explicitly
Objectives defined	Explicitly	Implicitly
Attributes defined	Implicitly	Explicitly
Number of alternatives	Infinite (large)	Finite (small)
Decision maker's control	Significant	Limited
Decision modelling paradigm	Process-oriented	Outcome-oriented
Relevant to	Design/search	Evaluation/choice

The second, more generally accepted manner in which MCDA methods may be classified, involves the use of the following three categories suggested by Belton and Stewart [17, p. 9]:

Value measurement models. In order to represent the extent to which one alternative may be preferred to another, numerical scores are determined. A score is determined for each criterion, whereafter these scores are combined so as to affect aggregation into higher-level preference models. It is thus possible to link a real number to each of the alternatives so as to produce a preference order of the alternatives.

Goal, aspiration or reference level models. Levels of satisfaction performance for each criterion are first determined. Thereafter, those alternatives that are best capable of achieving the desired levels of performance are sought.

Outranking models. Different possible actions are compared in a pairwise manner so as to determine the extent to which the selection or preference of one action over another may be affirmed. By combining the information thus collected on the preferences, for all criteria involved, outranking models determine the strength of evidence favouring one alternative above another. Outranking models are similar to value measurement models in that the alternatives are given a ranking, but outranking models do, however, not produce a value function indicating the extent to which one alternative is worse than its predecessor or better than its successor in the resulting ranking.

4.3.2 MCDA problem formulation

According to Amor *et al.* [9], Karami [100, p. 6], and Mota *et al.* [157, p. 28], an MCDA problem may be formulated as follows. Let $\mathcal{A} = \{a_1, \ldots, a_i, \ldots, a_n\}$ denote the set of possible alternatives when analysing a discrete decision space. A decision maker may then select a certain number of these alternatives in a specific manner from the alternatives or possible actions that are available to them. Let $\mathcal{G} = \{g_1, \ldots, g_j, \ldots, g_u\}$ denote the set of criteria (or attributes) relevant to the decision at hand. Each of the u criteria represents a dimension in which an alternative may be evaluated. A score e_{ij} may be assigned by the decision maker to alternative i according to criterion j, which is indicative of how well that alternative performs in the context of criterion j. These scores together form an $n \times u$ *evaluation table*, *decision table*, or *performance matrix*

$$\boldsymbol{E} = \begin{array}{c} \\ a_1 \\ \vdots \\ a_i \\ \vdots \\ a_n \end{array} \begin{array}{c} \begin{array}{ccccc} g_1 & \cdots & g_j & \cdots & g_u \end{array} \\ \begin{pmatrix} e_{11} & \ldots & e_{1j} & \ldots & e_{1u} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_{i1} & \ldots & e_{ij} & \ldots & e_{iu} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e_{n1} & \ldots & e_{nj} & \ldots & e_{nu} \end{pmatrix} \end{array}. \tag{4.1}$$

Each of the scores e_{ij} describes the value of alternative a_i against criterion g_j. It is thus possible to specify an MCDA problem instance by producing the triple $(\mathcal{A}, \mathcal{G}, \boldsymbol{E})$. In addition, each criterion j may be assigned a weight or relative importance value denoted by w_j. These weights should add up to one, that is $\sum_{j=1}^{u} w_j = 1$.

4.3.3 Preference modelling

According to Guitouni and Martel [82, p. 505], the vast majority of MCDA methods utilise preferences selected by a decision maker. If a and b represent two alternatives, it is possible to define the following four preferences, expressed as binary relations:

Preference situation $(a\,\boldsymbol{P}\,b)$. The decision maker strictly and strongly prefers alternative a to alternative b.

Weak preference situation $(a\,\boldsymbol{Q}\,b)$. The decision maker again strictly prefers alternative a to alternative b, but only weakly (*i.e.* the decision lacks conviction).

Indifference situation $(a\,\boldsymbol{I}\,b)$. The decision maker is indifferent between alternatives a and b. (*i.e.* there is no evidence to suggest that a is preferred to b). Unlike the relation $\boldsymbol{Q}$, which hesitantly indicates that one is preferred above the other, the relation $\boldsymbol{I}$ clearly states that the two alternatives are equivalent.

Incomparability situation $(a\,\boldsymbol{R}\,b)$. Alternative a and b are incomparable, indicating that the decision maker was not able to or refused to take a side, possibly due to a lack of information.

Colson and de Bruyn [43, p. 1205] explain how these four elementary relations may be combined using logical operators, to form *preference structures*. An example of such a preference structure is $a\,(\boldsymbol{P} \cup \boldsymbol{I})\,b$, which implies preference or indifference between alternatives a and b. This means that if $\boldsymbol{Q}$ is unavailable, one may select $\boldsymbol{P}$ or $\boldsymbol{I}$. Another example of a preference structure is, $a\,(\boldsymbol{P} \cup \boldsymbol{Q} \cup \boldsymbol{I})\,b$, or $a\,\boldsymbol{S}\,b$, which represents the *outranking relation*. It indicates the situation where strong evidence exists that alternative a is at least as good as b, with no evidence that would strongly support the opposite, namely that a outranks b [43, p. 1205], [82, p. 505].

4.3.4 Available MCDA methods

Although other methods also exist, the focus in this thesis falls on a very detailed list and description of many well-known MCDA methods presented by Guitouni and Martel [82, p. 508]. Guitouni and Martel categorised the family of MCDA methods into four categories, namely *mixed methods*, *elementary methods*, *single synthesizing criterion methods*, and *outranking methods*.

The first of the four categories, mixed methods, are those methods which cannot be categorised into any of the other three categories to be explained next. These methods are summarised in Table 4.4.

Linkov *et al.* [124, p. 19] state that elementary methods are those which utilise far less complex workings to produce results than more complicated methods, and may in many cases be carried out by hand. While multiple criteria may be present, the different criteria do not require the calculation of weights. For problems involving few alternatives and criteria, the best suited methods are elementary methods. Such problems are unfortunately not often a representative requirement of real-world scenarios. These methods are summarised in Table 4.5.

Single synthesising criterion methods, according to Akbulut [5, p. 22], are considered the more traditional approaches. As the name implies, these methods attempt to reduce the different criteria into a single index by utilising algorithms and rules. Each of the alternatives is evaluated independently. This class of methods is summarised in Table 4.6.

Martel and Matarazzo [139, p. 198] explain that outranking methods operate by comparing all alternatives by pair-building some form of binary relationship, after which relationships are examined between all the criteria in such a manner that final recommendations may be made. These methods are summarised in Table 4.7.

TABLE 4.4: *Two mixed MCDA methods.*

MCDA methods: Mixed methods	
QUALIFLEX	Consecutive mutations are utilised to produce a ranking of alternatives that support the qualitative data.
The Martel and Zaras method	Pairwise comparisons are performed using stochastic dominance; these are then used as partial preferences. It is then possible to use both a concordance index and a discordance index to produce an outranking relation.

Table 4.5: *Various elementary MCDA methods, as presented by Guitouni and Martel [82, p. 508].*

MCDA methods: Elementary methods	
Weighted sum	Possibly the best known MCDA method. The score of each alternative against all the criteria are summed in a weighted manner to determine a final performance score for each alternative. The alternatives are then evaluated based on their final performance scores.
Lexicographic method	Functions on the idea that some criteria are much more important than others. It evaluates all the alternatives against the most important criterion and systematically removes alternatives. Each next best criterion is used once the current criterion cannot refine the alternative list any further.
Conjunctive method	A minimal acceptable level for each criterion is defined. Those alternatives which do not satisfy the minimum level for all of the criteria are removed.
Disjunctive method	A desirable level for each criterion is defined. Those alternatives which do not equal or exceed the desirable level for any single criterion are removed.
Maximin method	A global performance score is associated with each alternative, based on its weakest or poorest evaluation.

Table 4.6: *Various single synthesising criterion MCDA methods, as presented by Guitouni and Martel [82, p. 508].*

MCDA methods: Single synthesising criterion methods	
TOPSIS (technique for order by similarity to ideal solution)	The best alternatives are those which are located closest to the ideal and farthest from the non-ideal solutions.
MAVT (multi-attribute value theory)	The values obtained from partial value functions of each criterion are aggregated to produce a global value function. The aggregation may occur in either a multiplicative, an additive or a mixed manner.
UTA (utility theory additive)	Ordinal regression is used to determine the value function for each criterion, whereafter the global value function is achieved in an additive manner.
SMART (simple multi-attribute rating technique)	Implements multi-attribute utility theory, but by using the weighted linear averages. This produces a very close approximation to utility functions. Examples of improvement of this method include SMARTS and SMARTER.
MAUT (multi-attribute utility theory)	A global utility function is determined by combining the values obtained through partial utility functions of each criterion. The aggregation may occur in either an additive, multiplicative or distributional manner.
AHP (analytic hierarchy process)	The problem is broken down into a more easily understood hierarchy of sub-problems. Thereafter, the various elements of the hierarchy are evaluated in a pairwise manner, the results of which are expressed in numeric terms. Finally, priorities are determined for each alternative.

EVAMIX	A method able to consider criteria based on qualitative or quantitative data. A dominance index is determined for each separately, which is then combined so as to present the dominance of each pair of alternatives.
Fuzzy weighted sum	The α-cut technique is utilised, as the α level sets assist in determining fuzzy utilities which are modelled on the simple additive weighted method.
Fuzzy maximin	Based on the same fundamental concepts as the Maximin method explain above, fuzzy numbers are produced from the assessment of the alternatives.

Table 4.7: *Various outranking MCDA methods, as presented by Guitouni and Martel [82, p. 508].*

MCDA methods: Outranking methods	
ELECTRE I	The first of the outranking methods, utilising the outranking relationship approach, functioning on the idea that an alternative may be removed if it is dominated to a certain extent.
ELECTRE IS	Identical to ELECTRE I, except that indifference thresholds are also introduced and utilised.
ELECTRE II	The opinion of the decision maker is converted into an ordinal scale and, as in the other ELECTRE methods, agreement and disagreement indices are utilised.
ELECTRE III	A ranking procedure in which a credibility index is utilised so as to express the outranking.
ELECTRE IV	Similar to ELECTRE III, except that criteria weights are not utilised.
ELECTRE TRI	Similar to ELECTRE III, but utilised for sorting. The different alternatives are sorted into the ordered categories by use of conjunctive and disjunctive techniques.
PROMETHEE I	An outranking method in which the decision maker's preference may be described using one of six criteria functions. It utilises both entering and leaving flows and produces a partial ranking of alternatives.
PROMETHEE II	Building on the principles of PROMETHEE I, a complete ranking is produced instead of a partial ranking of the alternatives. The entering and leaving flows are aggregated as opposed to being considered separately.
MELCHIOR	An extension of ELECTRE IV.
ORESTE	The only required input of this method is an importance ranking of the criteria and qualitative evaluations of the alternatives.
REGIME	A pairwise comparison matrix is constructed, populated with one of three scores. If one alternative dominates another, a value +1 is assigned; if the pair of alternatives are equal then a value 0 is used. Finally, if negative-dominance is present, a value −1 is assigned. By combining these scores the alternatives may be ranked in a complete pre-order.
NAIADE (novel approach to imprecise assessment and decision environments)	Distance semantics operators are utilised to evaluate the pairs of alternatives, whereafter entering and leaving flows are determined (similar to PROMETHEE). Probability distributions are finally produced by transforming the fuzzy evaluations.

A detailed discussion on ELECTRE III now follows, because it is the MCDA method applied later in this thesis.

4.3.5 ELECTRE III

The class of *ELimination Et Choix Traduisant la REalite* (ELECTRE) methods is a family of outranking MCDA methods. The first member of the family, ELECTRE I, was initially proposed in 1965; the other methods have since joined the family. A very popular and widely employed member of this family, known as ELECTRE III, is described in some detail in this section [130, p. 1].

Majdi [130, p. 39] states that ELECTRE III consists of two main phases: *construction of the outranking relation* and *exploitation of the outranking relation*. The concepts of *preference*, *concordance*, *non-discordance*, and *credibility* are incorporated during the first phase and are described first, while the elements of *distillation* and *final ranking* occur during the second phase and are explained thereafter.

Types of preference criteria

According to Colson and de Bruyn [43, p. 1206], a *preference criterion* is a decision rule based on a specific ranking of two alternatives. There are two classes of preference criteria, namely *true criteria* and *pseudo-criteria*.

True criteria utilise the two preference situations $\boldsymbol{P}$ and $\boldsymbol{I}$ described in §4.3.3, and thus assign a preference if any difference is present on the ranking scale, with indifference only being assigned when the two alternatives are considered equal by the decision maker.

To compensate for the limited discriminant power of humans, the class of pseudo-criteria has been introduced. In pseudo-criteria, the outranking relation $\boldsymbol{S}$, which utilises the three preference situations $\boldsymbol{P}$, $\boldsymbol{Q}$, and $\boldsymbol{I}$ of §4.3.3, is combined with two discriminant thresholds, known as the *preference threshold* and *indifference threshold*, denoted by p and q, respectively [5, p. 27]. The preference threshold allows for the separation of the strong preference $\boldsymbol{P}$ from the weak preference $\boldsymbol{Q}$, whereas the indifference threshold partitions the weak preference $\boldsymbol{Q}$ from the indifferent preference $\boldsymbol{I}$ [43, p. 1206]. In the special case of pseudo-criteria where $p = q = 0$, a true criterion is obtained [43, p. 1207].

Concordance and Non-discordance

Akbulut [5, p. 26] and Figueira *et al.* [66, p. 159] explain that outranking models rely on the concepts of *concordance* and *non-discordance*, since by these notions it is possible to test the outranking relation $\boldsymbol{S}$ for acceptance. Note that $a\ \boldsymbol{S}\ b$ implies that "a outranks b." The principle of concordance requires that a majority of the j criteria, with their relative weightings taken into consideration, should be in favour of the relation $a\ \boldsymbol{S}\ b$ in order for it to be valid. The second principle, non-discordance, states that $a\ \boldsymbol{S}\ b$ may be classified as valid provided that none of the criteria which form part of the minority strongly oppose the assertion.

Credibility index

The outranking relation depends on the definition of a *credibility index*, which may be used to quantify the credibility of the allegation $a\ \boldsymbol{S}\ b$ [205, p. 20]. The credibility index $S(a, b)$ comprises two other indices, namely a concordance index $c_j(a, b)$ and a discordance index $d_j(a, b)$ for each criterion j.

The definition of two parameters is now required. The first is the notion of *importance coefficients*, which are the criterion weights w_j previously mentioned in §4.3.2. In the case of the ELECTRE methods, these weights indicate the voting power or weight that each criterion j carries to be in favour of a specific alternative. Let $g_j(a)$ denote the evaluation of alternative a on criterion j. The second type of parameter is the *veto threshold*, denoted by v_j. This parameter indicates the power that criterion j has to impose its veto power, *i.e.* the power that criterion j has to be against an allegation of the form "a outranks b," in the case that the difference between $g_j(a)$ and $g_j(b)$ is greater than the threshold. The outcome may, for example, be "a may not outrank b" since the evaluation $g_j(b)$ of alternative b according to criterion j outperforms that of $g_j(a)$ by an amount greater than the specified veto threshold. The indication that criterion j has imposed its veto power is that the discordance index $d_j(a\ \boldsymbol{S}\ b)$ equals one for that specific criterion [130, p. 19].

By combining all of the above knowledge, it is now possible to define the overall concordance index as

$$C(a,b)=\frac{1}{W}\sum_{j=1}^{u} w_j c_j(a,b)$$

[5, p. 28], where

$$W=\sum_{j=1}^{u} w_j,$$

and the concordance index for each criterion as

$$c_j(a,b)=\begin{cases} 1, & \text{if} \quad g_j(a)+q_j \geq g_j(b),\\ 0, & \text{if} \quad g_j(a)+p_j \leq g_j(b),\\ \frac{p_j+g_j(a)-g_j(b)}{p_j-q_j}, & \text{otherwise,} \end{cases}$$

where p_j and q_j denote respectively the preference threshold and indifference threshold for criterion j, as described above. The second index, the discordance index, is defined as

$$d_j(a,b)=\begin{cases} 0, & \text{if} \quad g_j(a)+p_j \geq g_j(b),\\ 1, & \text{if} \quad g_j(a)+v_j \leq g_j(b),\\ \frac{g_j(b)+g_j(a)-p_j}{v_j-p_j}, & \text{otherwise,} \end{cases}$$

where p_j and v_j denote respectively the preference threshold and veto threshold of criterion j [130, p. 43]. Notice that the discordance index equals one when criterion j exercises its veto power. By combining the above two indices, the credibility index

$$S(a,b)=\begin{cases} C(a,b), & \text{if} \quad d_j(a,b)\leq C(a,b) \text{ for all } j,\\ C(a,b)\ \prod_{j\in J(a,b)} \frac{1-d_j(a,b)}{1-C(a,b)}, & \text{otherwise,} \end{cases}$$

is obtained for the pair of alternatives (a,b), where $J(a,b)$ represents the set of criteria for which $d_j(a,b) > C(a,b)$ for all j. In the case that no veto threshold v_j is specified, $d_j(a,b)$ equals zero for all pairs of alternatives, and subsequently $S(a,b)=C(a,b)$ for all pairs of alternatives.

To summarise, the input required from the decision maker for an ELECTRE method is the three sets of parameters $p_1,\ldots,p_u, q_1,\ldots,q_u$, and $v_1,\ldots,v_u$ for u criteria, the u criterion weights $w_1,\ldots,w_u$, and an indication of whether each of the u criteria must be maximised or minimised.

Distillation and final ranking

Distillation, the first part of the second phase of ELECTRE III, involves exploitation and is an automated procedure which may be utilised to rank alternatives. The process of distillation occurs in three steps, namely *descending distillation*, *ascending distillation*, and *combining distillations* [130, p. 44].

Descending distillation produces a first pre-order, Z_1. According to descending distillation, the best alternatives are considered first and the worst last. This is achieved by determining a *qualification* of each alternative. The qualification of alternative a is denoted by $Q(a)$ and is calculated as the number of alternatives that a outranks, less the number which outrank a. By considering the qualifications of all the alternatives, those with the single highest qualification are grouped and labelled as *distillation one*, denoted by D_1. It is possible for more than one alternative to belong to a single distillation level. The alternatives belonging to D_1 are removed from the original data set and the step is repeated for the alternatives that remain in order to produce the second distillation, denoted by D_2. This process is performed iteratively until all the alternatives have been assigned distillation levels [205, p. 22].

According to ascending distillation, denoted by Z_2, a second pre-order is formed in the same manner as in descending distillation until all the qualifications have been assigned to alternatives, but thereafter the distillation process is performed by considering those alternatives with the lowest qualifications first.

According to Shofade [205, p. 23], once both the above pre-orders Z_1 and Z_2 have been formed, they may be combined, based on the intersection of the pre-orders, so as to determine a final ranking of alternatives. The final ranking is a partial pre-order (or partial ranking), meaning that the ranking of any two alternatives may be in one of three forms. An alternative may either be preferred to another, indifferent to another, or incomparable with another. The situation of incomparability occurs if Z_1 ranks the one above the other, and Z_2 results in the opposite.

Selecting indifference, preference, and veto thresholds

According to Rogers and Bruen [188, p. 548] the relationship between the indifference (q), preference (p), and veto (v) thresholds should be of the type $q \leq p \leq v$. They emphasise that the original designer of ELECTRE III, Roy [190], stated that space should be provided for an element of subjectivity with regards to selecting p and q. The selection of these thresholds depends strongly on the specific criterion and available knowledge of the field. The following example illustrates this. Rogers and Bruen [188, p. 547] conducted research on ranking alternatives of various projects and elements affecting their environmental impact assessments. One of the criteria used in the study was the noise (measured in dB(A)) to be produced by each project. They found that q should be set at the point for which one alternative is measurably distinguishable from another. This point was found to be between two and three dB(A).

Specific guidelines of the nature presented in the example above that relate to the thresholds of criteria might well not exist for the criteria or within the field in which the decision maker finds himself. In such a case, at the very least, the relationship between thresholds of $q \leq p \leq v$ should be respected. Some other guidelines provided by Roy [190] may also be utilised. The first is that p should be substantially greater than q — this value may be more than double the value of q. Despite the difficulty in selecting the 'correct' values for and relationship between p and q, Bouyssou [25, p. 63] states that this guideline should be preferred to using only 'true' criteria (*i.e.* $p = q = 0$), as described above. The second guideline refers to the relationship between v and p. Although v may be equal to p, it is more general for it to be evidently greater than p — even three, five or ten times more. In addition, the distance between v and p should decrease as the importance weighting of that criterion increases, and *vice versa*. The logic behind this is that the less important a criterion is deemed, the less effect the veto threshold of that criterion will have on the overall ranking [190].

4.3.6 A framework for selecting MCDA methods

According to Guitouni and Martel [82], no single MCDA method should be considered best for all decision making situations. For this reason, many attempts have been made in the literature to develop a framework for narrowing down the initially large list of available MCDA methods to a single suitable method or subset of MCDA methods which may be used to solve a given decision making situation [157, p. 12]. One such framework is attributed to Hwang and Yoon [95], who utilised a tree diagram to assist the decision maker in identifying a suitable MCDA method or subset of MCDA methods. Another is due to Kornyshova and Camille [112], who developed four facets for assisting the decision maker to select a suitable MCDA method, by analysing the characteristics of the decision making situation with respect to the four facets. A third framework, and the one detailed in this section, was developed by Guitouni and Martel [82].

The framework of Guitouni and Martel [82, p. 512] starts by presenting the following seven guidelines which may be used to asses the current decision making situation and aid the decision maker during the process of selecting an appropriate MCDA method:

1. Identify the decision maker and the *operational approach* that is in line with his or her expectations.
2. Identify the decision maker's way of thinking and select an appropriate preference model.
3. Identify an appropriate decision problematic[1] in line with decision maker's expectations.
4. The selected MCDA method should be able to handle the input information.
5. The degree of compensation, discrimination power, and inter-criteria information of the MCDA method must be taken into consideration.
6. Ensure that the fundamental hypothesis of the method is verified.
7. Consider whether there exists software packages and tools to apply and support the selected MCDA method.

Of the above-mentioned seven guidelines, guidelines 1, 3, and 4 are discussed most in the literature, because it is widely believed that selecting MCDA methods based on these three guidelines may reduce the list of suitable methods down to a manageable size. These three guidelines are also not complex in nature and are thus easy to apply in a framework. For this reason only the inner workings of guidelines 1, 3, and 4 are discussed below and considered in the remainder of this thesis [156, p. 17].

According to Roy [190], it is possible to partition decision making situations into four different problematics. These four problematics are the ways in which MCDA may provide decision support to a decision maker, and are described below.

Description problematic ($P.\delta$). This problematic is not concerned with making recommendations, but rather with presenting various alternatives and their related consequences, thus describing a suitable set of alternatives and criteria. Belton and Stewart [17, p. 15] view this problematic as one in which the decision maker seeks to understand what is and is not possible.

[1] A problematic refers to the expected outcome of a decision or the manner in which a problem is presented.

Choice problematic ($P.\alpha$). The aim in this problematic is to find either a single or restricted set of alternatives that meet the predefined criteria.

Sorting problematic ($P.\beta$). In a sorting (also known as *screening*) problematic, each alternative is assigned to a predefined category.

Ranking problematic ($P.\gamma$). The alternatives are ranked or ordered after being compared with each other in the ranking problematic.

Since the focus of this thesis is on ranking, a further brief explanation of the ranking problematic is included. According to Ehrgott *et al.* [61, p. 299], ranking is the process of assigning a rank or label of 1 to the best alternative, a label of 2 to the second best, and so forth. In the case of a tie between two or more alternatives, all the alternatives involved in the tie are assigned the same label. For this reason the term *rank levels* is often preferred rather than *ranks*.

The three main types of data which may serve as input data to MCDA methods, according to Guitouni and Martel [82, p. 511], are qualitative data, quantitative data, or mixed data. Mixed data simply refers to the merging of qualitative and quantitative data. The reader is referred back to §2.1 for a more detailed discussion on the differences between qualitative and quantitative data.

A further manner in which MCDA methods may be categorised is by classifying them as either *deterministic* or *non-deterministic*. According to Gil-Aluja [78, p. 210], the early MCDA methods were all deterministic, but advances have since been made to include methods which may be used for non-deterministic analyses. Fuzzy MCDA methods are examples of methods that utilise non-deterministic information features.

Each of the twenty eight MCDA methods reviewed briefly in §4.3.4 are presented again in Table 4.8 together with an indication of how they relate to guidelines 1, 3, and 4 of Guitouni and Martel [82, p. 512].

Since Guitouni and Martel [82, p. 512] presented the seven guidelines described above, subsequent work has been done in other studies to combine these guidelines into a typological tree, so as to make it easier for a decision maker to determine a single MCDA method or subset of MCDA methods that are considered most appropriate for the decision in question. In particular, Mota [156, p. 15] has expanded on the work of others and has presented a typological tree involving four stages. The tree function at each level or stage represents a question and, based on the decision maker's answer to this question, subsets of MCDA methods are eliminated. This is repeated four times until only a select group of MCDA methods remain which, according to the typology tree, are best suited for the decision making situation for which answers were provided. As mentioned earlier, guidelines 1, 3, and 4 of Guitouni and Martel [82, p. 512] are discussed most in the literature. This is also the case in the typology tree proposed by Mota [156, p. 15], since the foundations of the four stages of the typology tree are based on these guidelines. In the case where one would like to apply the remaining guidelines, this may be done after the four-stage typology tree has been implemented in order to refine the result even further, but only if required [156, p. 17].

In stage one of Mota's typological tree, presented schematically in Figure 4.2 (which is linked to guideline 1), the decision maker is presented with the question "*What is the operational approach?*" The four possible answers correspond to the four categories in which the twenty eight methods were presented in §4.3.4, namely *elementary approaches*, *single synthesizing criterion approaches*, *outranking synthesizing approaches*, and *mixed approaches*. Mota [156, p. 17], however, does not include *elementary approaches* as a possible answer to the question of stage one — only the remaining three options are included. A fourth possible answer he includes is

Table 4.8: *The MCDA methods of Tables 4.4–4.7 in the context of the MCDA selection guidelines put forward by Guitouni and Martel [82, p. 515].*

MCDA method	Guideline 3: Problematic	Guideline 4.1: Kind of information			Guideline 4.2: Information features	
		Qualitative	Quantitative	Mixed	Deter.	Non-deter.
Mixed methods						
QUALIFLEX	γ	✓			✓	
Martel and Zaras method	γ	✓	✓	✓		✓
Elementary methods						
Weighted sum	α		✓		✓	
Lexicographic method	α	✓	✓	✓	✓	
Conjunctive method	N.A.	✓	✓	✓	✓	
Disjunctive method	N.A.	✓	✓	✓	✓	
Maximin method	α	✓	✓		✓	
Single synthesizing criterion						
TOPSIS	α		✓		✓	
MAVT	α		✓		✓	
UTA	α	✓			✓	
SMART	α		✓		✓	
MAUT	α		✓			✓
AHP	α, γ		✓		✓	✓
EVAMIX	α, γ	✓	✓	✓	✓	
Fuzzy weighted sum	α	✓	✓	✓		✓
Fuzzy maximin	α	✓	✓		✓	✓
Outranking methods						
ELECTRE I	α	✓	✓	✓	✓	
ELECTRE IS	α	✓	✓	✓	✓	
ELECTRE II	γ	✓	✓	✓	✓	
ELECTRE III	γ	✓	✓	✓	✓	
ELECTRE IV	γ	✓	✓	✓	✓	
ELECTRE TRI	β	✓	✓	✓	✓	
PROMETHEE I	γ	✓	✓	✓	✓	
PROMETHEE II	γ	✓	✓	✓	✓	
MELCHIOR	γ	✓			✓	
ORESTE	γ	✓			✓	
REGIME	γ	✓			✓	
NAIADE	γ	✓	✓	✓	✓	✓

interactive approaches. Interactive approaches, according to Teghem *et al.* [223, p. 1315], involve the decision maker in an iterative manner. Preferences are requested from the decision maker, the answers of which are re-examined by the method, whereafter the decision maker is again presented with other possible preferences. This process is then repeated until the list of possible alternatives has been narrowed down to an acceptable number, as specified by the decision maker. Such methods do not function according to traditional rules; preferences are instead discovered during the iterative process. Four examples of such interactive approaches are STRANGE, STEM STRANGE, STEM RBSTRANGE, MOMIX. By selecting a family of methods during this stage, the number of remaining methods is largely refined.

The next two stages of the typological tree relate to guideline 4 and thus to the data involved in the decision process. Stage two, illustrated in Figure 4.3, poses the question "What kind of information is considered?" The possible answers to this question were discussed above. Stage three of Mota's typological tree, illustrated in Figure 4.4, is focused on the nature of the information and presents the question "What is the nature of the information?" Descriptions of the three possible answers may again be found in the discussion above.

The final and fourth stage of the typological tree, illustrated in Figure 4.5, is associated with guideline 3 and presents the question "Which decision problematic is addressed?" After the decision maker has followed the process laid out by the four stages of the typology tree, a subset of appropriate MCDA methods of manageable size should have been produced.

Due to the different but overlapping contexts discussed within this thesis, the terms *alternative*, *student*, *applicant*, and *observation* will henceforth be used interchangeably.

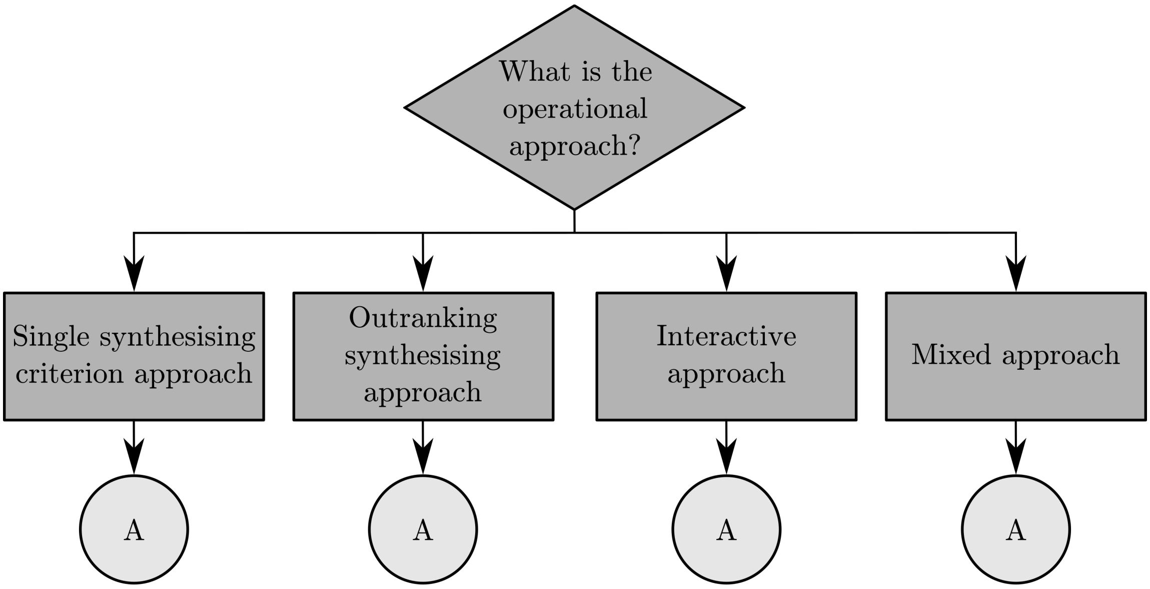

What is the operational approach?
Single synthesising criterion approach
Outranking synthesising approach
Interactive approach
Mixed approach
A
A
A
A

Figure 4.2: *The first stage of Mota's typological tree, adapted from Mota [156, p. 16].*

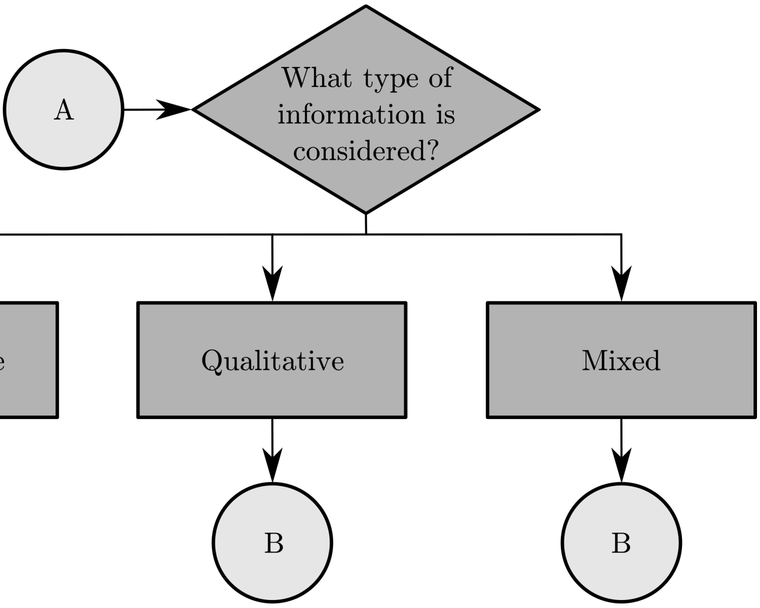


Figure 4.3: *The second stage of Mota's typological tree, adapted from Mota [156, p. 16].*

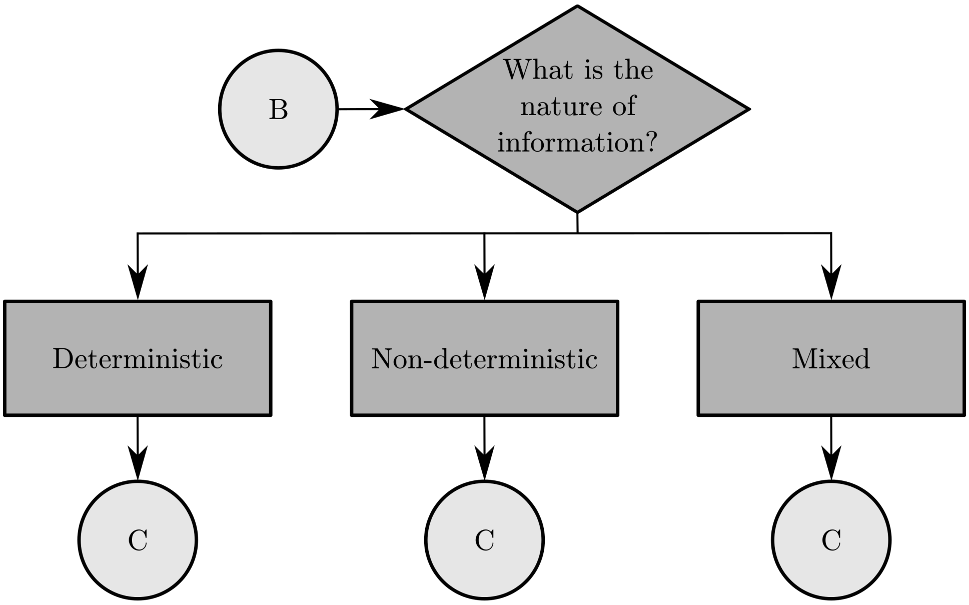


FIGURE 4.4: *The third stage of Mota's typological tree, adapted from Mota [156, p. 17].*

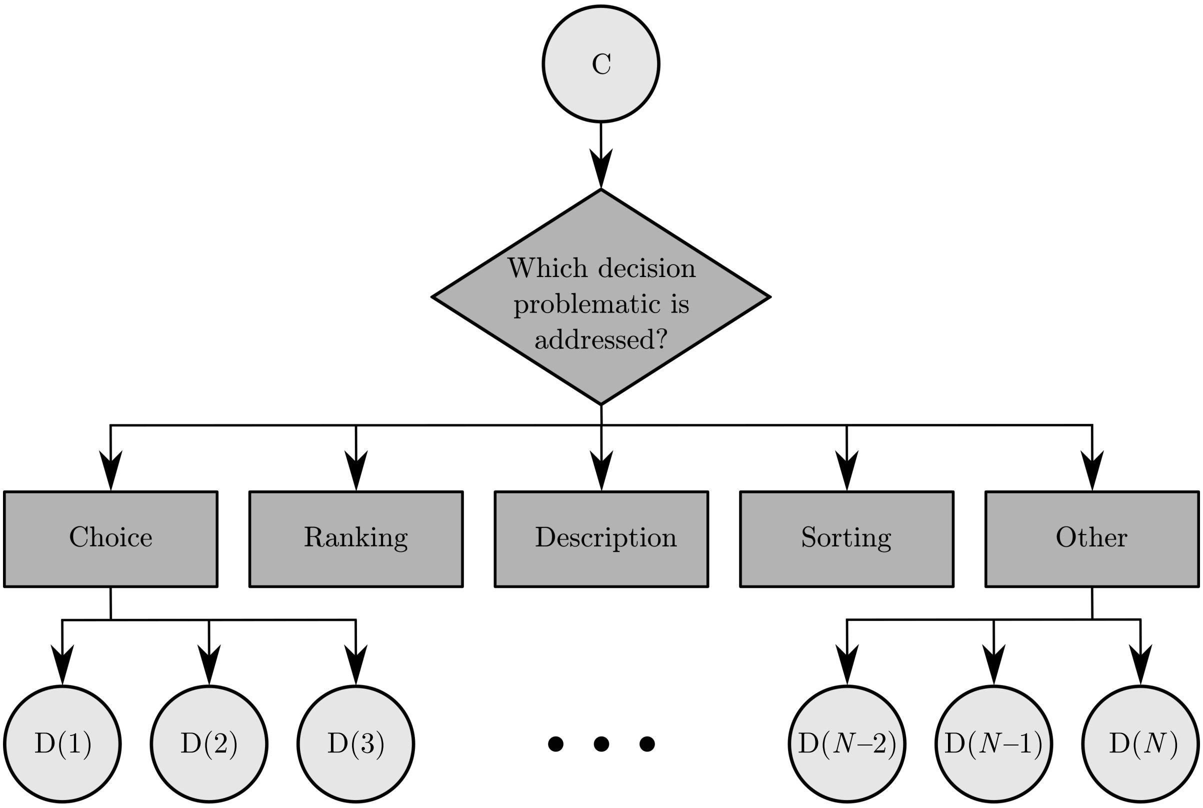


FIGURE 4.5: *The fourth stage of Mota's typological tree, adapted from Mota [156, p. 18].*

4.4 Chapter summary

This chapter was dedicated to a discussion on ensemble learning and MCDA. After having presented a background on ensemble learning, three mechanisms were reviewed in §4.1 by which an ensemble method may be constructed. The third type of mechanism was further expanded upon, because it is of particular relevance to the topic of this thesis. The two types of integration strategies used for combining base model predictions (*i.e. static integration* and *dynamic integration*) were described in §4.2.

The next section was devoted to MCDA. The first part of the section contained a description of the different manners in which MCDA methods may be categorised. Thereafter, a procedure for formulating a typical MCDA problem was presented. This was followed by a brief description of twenty eight well-known MCDA methods, and a more detailed discussion on ELECTRE III, which is of particular importance to the topic of this thesis. In the final part of the section a framework for selecting MCDA methods based on a specific decision situation was reviewed.

Part II

Decision support system

CHAPTER 5

Decision support system architecture

Contents

5.1 The proposed DSS framework 109
5.2 Data preparation 110
5.3 Model base 111
5.4 Chapter summary 113

In this chapter, an architecture is presented of the DSS framework proposed in this thesis for predicting output and ranking various alternatives using multiple prediction models. This is done by graphically illustrating the main information flows and components of the DSS. Two main subcomponents, namely those of data preparation and the model base, are then presented and discussed in more detail. The model base subcomponent is presented in considerable detail due to its importance within the framework. The chapter closes with a brief summary of the framework presented.

5.1 The proposed DSS framework

The three main components which make up the DSS proposed in this thesis for predicting the output class and ranking of various alternatives, as discussed in §2.3.2, are the database (C), the model base (M), and the user interface (G). The interaction between these three components and the flow of information between them, are illustarted diagrammatically in Figure 5.1.

Two main activities occur within the framework of Figure 5.1. First, the model base (M) requests past applicant data (J) so as to facilitate learning, indicated by the dashed lines. This is performed by requesting past applicant data (E) from the database (C), and then preparing the data (F). Data preparation involves cleaning the data by removing errors and ensuring that it is in the correct format. The cleaned, past applicant data may then be passed to the model base.

The second main activity is concerned with the input that the user may provide to the system (B). The user may request results (B.2) from the model base (M) in the form of predictions and/or rankings of students (B.2.2) upon providing the model base (M) with new applicant data (B.1). The results requested by the user (K) also contain information on the specific dependent and independent variables the user would like to consider. The first activity performed by the model base (M) is to validate the assumptions related to the data and models within the model base, indicated by the dotted line. The results of the assumption validation process are passed to the user (H) who considers the results (B.2.1) and is able select which models within

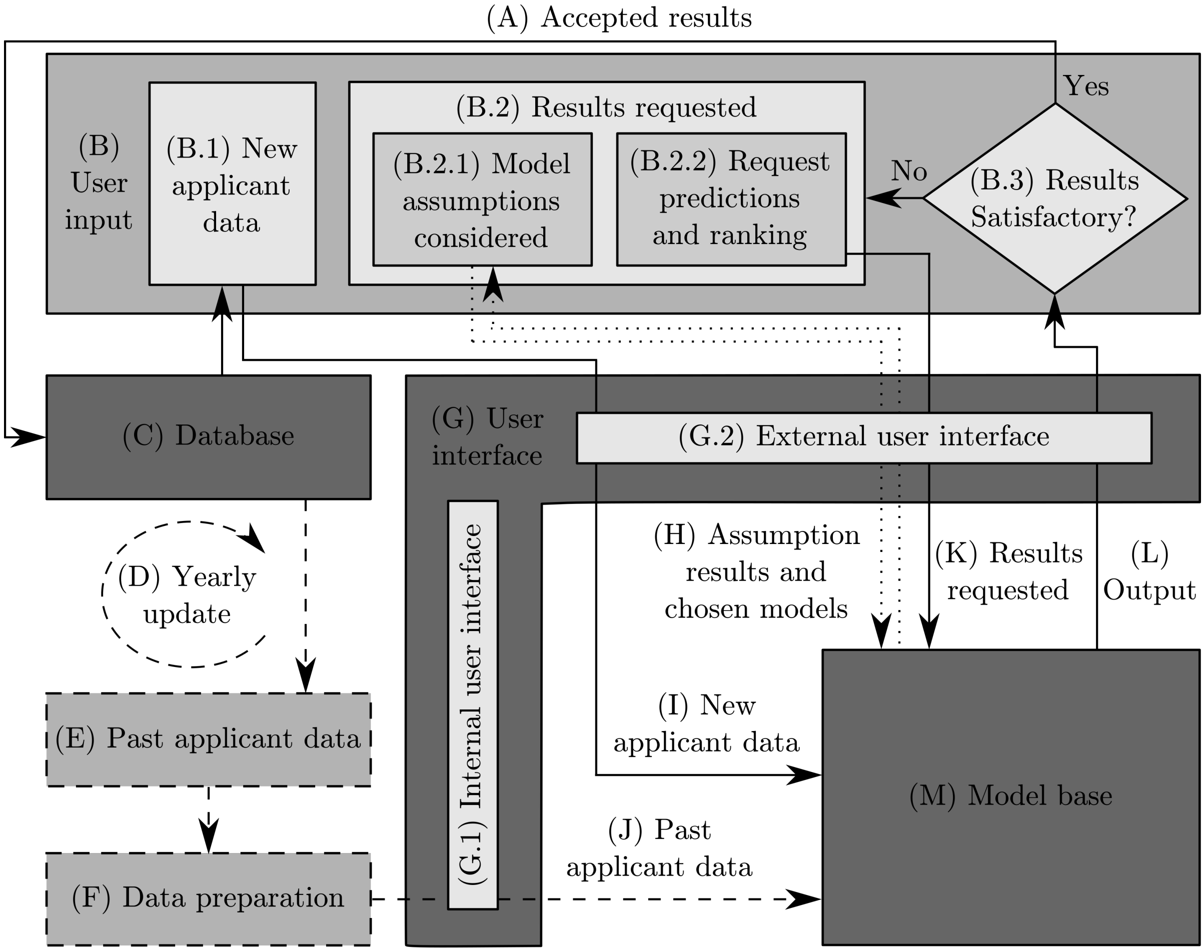


FIGURE 5.1: *Overall working of the proposed DSS.*

the model base he or she would still like to be activated, despite possible assumption violations that may exist for those models. The model base (M) is then passed the selected base models (H). Using the activated base models (H) and their training in the context of past applicant data (J), the model base (M) produces results for the new applicants (I) and presents them to the user (L). The user may then consider them (B.3) and, if found to be satisfactory, they may be accepted and passed (A) to the database (C). Alternatively, if found to be unsatisfactory, the user may request new results.

The two components of the DSS framework that require further discussion are the data preparation (F) and model base (M) components. These are described in more detail in the following sections.

5.2 Data preparation

Recall that in §2.2 a discussion took place on data preparation in terms of the four steps of data gathering (D), data integration (F.1), data extraction (F.2), and assessing data quality and performing data cleaning (F.3). The first step of data preparation, namely data gathering, should be performed on an annual basis and is represented by component (D) in Figure 5.1. The remainder of the data preparation takes place in the three sub-blocks within block (F) of Figure 5.2. The process within the data preparation block should be repeated each year, once data collection has taken place.

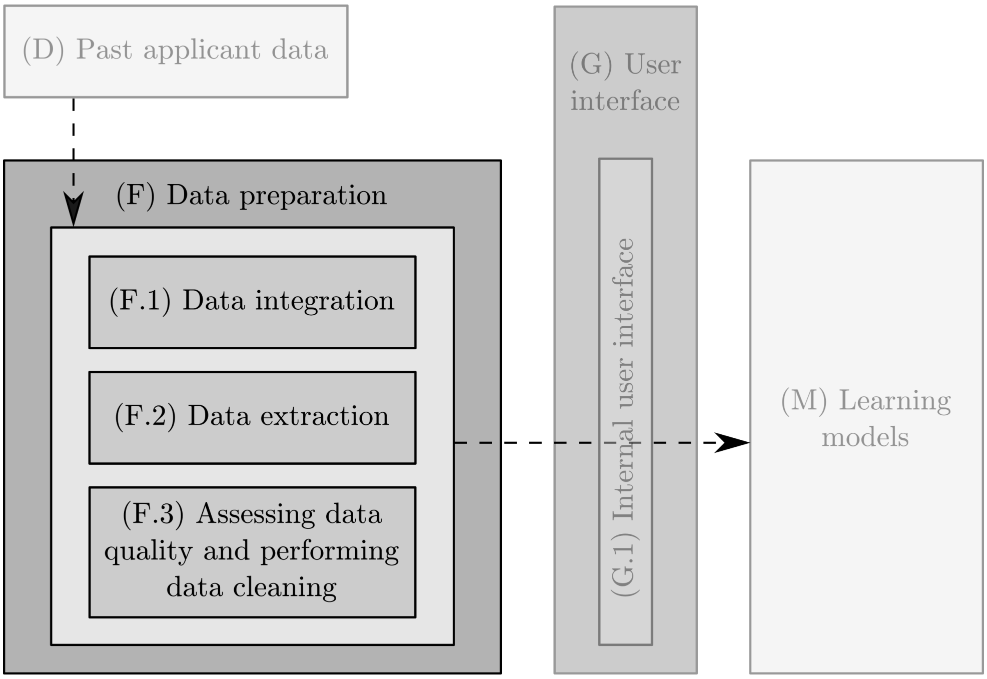


Figure 5.2: *Working of the data preparation component.*

The fourth data preparation step, namely assessing data quality and performing data cleaning (F.3), may further be broken down into three components involving the treatment of missing values (F.3.1), error values (F.3.2), and outlier and influential values (F.3.3), as illustrated in Figure 5.3. The details of the processes followed in these three components were discussed in §2.2.4. Note that outlier and influential observation detection is also a possible requirement for certain base models, and so this step may be performed during data preparation or only later during assumption analysis within the model base (M). The working of the model base is discussed next.

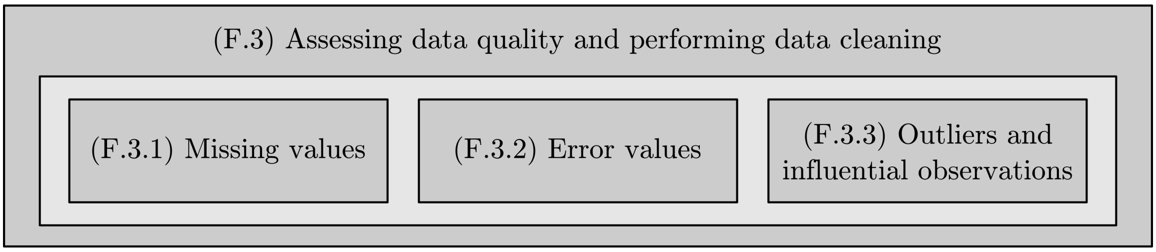

(F.3) Assessing data quality and performing data cleaning
(F.3.1) Missing values
(F.3.2) Error values
(F.3.3) Outliers and influential observations

Figure 5.3: *Working of the assessing data quality and performing data cleaning component.*

5.3 Model base

An example decision framework from the literature, which makes use of multiple techniques or models, was presented in Figure 2.4. This ensemble type framework inspired the design of the proposed model base component in Figure 5.4.

The information flows and working of the model base (M) proceed as follows. Recall from §5.1 that the user requests predictions and/or rankings of new applicants (B.2.2) by providing data on the new applicants (B.1) and selecting dependent and independent variables to be included in the analysis. All these preferences and data are accumulated in block (M.1), which also pulls in the cleaned past applicant data (J). The remainder of the activities of the model base are partitioned into the processes of learning the models (M.2), and predicting outcomes and ranking applicants (M.3).

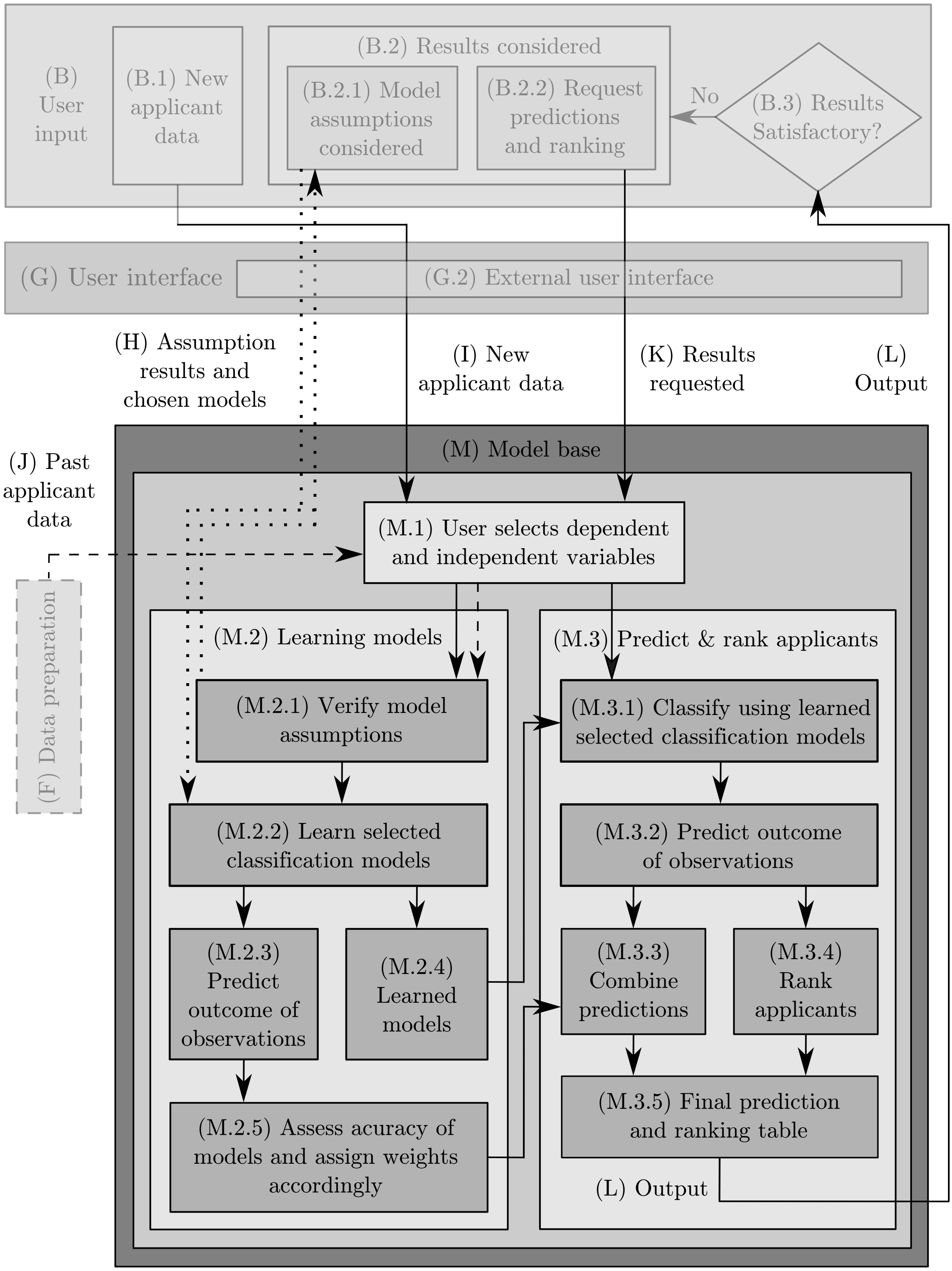

(B) User input
(B.1) New applicant data
(B.2) Results considered
(B.2.1) Model assumptions considered
(B.2.2) Request predictions and ranking
No
(B.3) Results Satisfactory?
(G) User interface
(G.2) External user interface
(H) Assumption results and chosen models
(I) New applicant data
(K) Results requested
(L) Output
(J) Past applicant data
(F) Data preparation
(M) Model base
(M.1) User selects dependent and independent variables
(M.2) Learning models
(M.2.1) Verify model assumptions
(M.2.2) Learn selected classification models
(M.2.3) Predict outcome of observations
(M.2.4) Learned models
(M.2.5) Assess acuracy of models and assign weights accordingly
(M.3) Predict & rank applicants
(M.3.1) Classify using learned selected classification models
(M.3.2) Predict outcome of observations
(M.3.3) Combine predictions
(M.3.4) Rank applicants
(M.3.5) Final prediction and ranking table
(L) Output

FIGURE 5.4: *Working of the model base component.*

First, the process of learning the models (M.2) involves twice evaluating the data and model assumptions and passing the results back to the user (H), who may then, after the second round, select the final models to be used. During the first stage, the assessment results related to any assumption for which the corrective steps may include removal of observations, such as outlier and influential value identification, should be presented to the user so that he or she may select an appropriate corrective action (H). During the second stage, the remainder of the assumptions are validated (M.2.1) and their results are passed to the user (H).

All of the past applicant data (J) may be used to learn the selected classification models (M.2.2). At this point the classification models may be considered learned (M.2.4). The outcome of the validation set of students may then be predicted (M.2.3). By assessing the predictive accuracy of each classification base model, static weights may be assigned to them.

The process of predicting the outcome and ranking of new applicants (M.3) occurs as follows. Using the learned models (M.2.4), new applicant data (M.1), selected dependent and independent variables (M.1), and selected classification base models (M.3.1), the outcomes of the new applicants are predicted (M.3.2). Thereafter, using the model weights (M.2.5), the predictions for the new applicants by the various classification models are combined so as to produce a single prediction for each applicant (M.3.3). This combination process may proceed according to one of the static integration strategies discussed in §4.2.1.

By studying the static combined weighting technique (illustrated in Table 4.1) and the general MCDA technique formulation of (4.1), it should be evident that many similarities exist between the formats of the two, so much so that an outranking MCDA method may be applied to rank the applicants into rank levels. This may be achieved by treating the applicants as the alternatives, the different classification models (M.3.1) as criteria, the predictions of the models as the criterion scores, and the model weightings (M.2.5) as the criteria weights. It is then possible to rank the new applicants (M.3.4) in an appropriate manner so as to produce rank levels for each. This ranking is performed according to one of the applicable MCDA outranking methods identified during the MCDA method selection process described in §4.3.6. A final table containing both the prediction and ranking of each new applicant may then be formed and produced as output (L) to be considered by the user (B.3).

Examples of classification models that may be used within the proposed framework as base models include those discussed in §3. A list of all base models are made available to the user to select from once an assumptions analysis has been carried out.

5.4 Chapter summary

An architecture was presented in this chapter for the DSS framework proposed in this thesis for predicting the output class and ranking of various alternatives using multiple models. First a holistic view was taken after which the focus shifted to separate components of the DSS, describing the information flows between them. The two components involving data preparation and model base subcomponents were expanded upon with a particular focus on the model base component due to its importance within the framework.

Throughout the explanation of the framework, reference was made to the applicable literature review sections which form the basis of the various components. In addition, the generic ability of the framework was emphasised by pointing out that base models, prediction weighting method, and applicant ranking method are not fixed, but may be preselected by the user of the system.

CHAPTER 6

Decision support system development

Contents

6.1 System development methodology and modelling tools 115
6.2 Study objectives and user requirements 116
6.3 Analysis of the system needs 117
6.4 Design of the system 120
6.5 Building of the system 120
6.6 Quality assurance, verification and validation 137
6.7 Chapter summary 142

The aim in this chapter is to describe the development and working of a concept demonstrator of the DSS design of Chapter 5. It is first described how various systems development methodologies and modelling tools are applied on top of the SDLC foundation so as to develop the concept demonstrator. Next, the user requirements are presented, referring back to the study objectives of §1.5. An analysis of the system needs then follows, focussing on the current state of the main industry partner's bursary application system. Next, a discussion takes place on the design of the system, previously presented in Chapter 5. This is followed by a section devoted to a presentation of the building of the system, by specifically discussing the data preparation, data assumption considerations, software considerations, selection of the framework elements, and GUI considerations. The final part of that section contains a system walk-through of the proposed DSS GUI using actual data. Next, a section is devoted to quality assurance, verification, validation, and implementation considerations of the DSS. The chapter concludes with a brief summary of the topics covered.

6.1 System development methodology and modelling tools

The DSS concept demonstrator development process utilises the seven phases of the SDLC (reproduced in Figure 6.1 and previously discussed in §2.3.3) as its foundation. The seven phases are performed in the context of the systems development methodologies selected and specific modelling tools are employed during selected phases so as to assist in their application.

The first few phases of the SDLC are performed according to the waterfall methodology (described in §2.3.4), with its strong focus on structure and documentation. The OOM methodology then becomes more prominent during phase three (analysing the system needs) of the SDLC, as the UML use case and activity diagrams are used to visualise how the system interacts with its environment based on the user requirements.

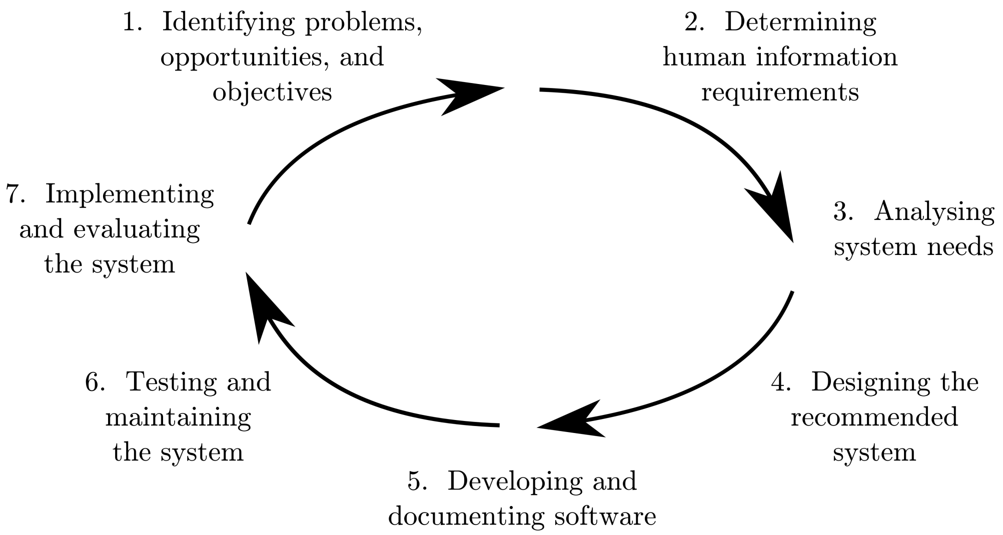

1. Identifying problems, opportunities, and objectives
2. Determining human information requirements
3. Analysing system needs
4. Designing the recommended system
5. Developing and documenting software
6. Testing and maintaining the system
7. Implementing and evaluating the system

Figure 6.1: *The seven phases of the systems development life cycle (SDLC), adapted from Kendall and Kendall [103, p. 36].*

Phase four (system design) is again conducted according to the waterfall methodology, meaning that emphasis is placed on thorough documentation. Phase four relates to the design of the DSS architecture.

The remainder of the phases are carried out in an agile manner, which implies that the development process occurs at a high speed and obtain feedback from the client for consideration at more than one point during the design process. This allows for valuable changes to be made to the system before its conclusion.

The physical building of the system commences in phase five (software development and documentation). It is described in phase six how continuous testing and verification was applied, as required by the agile methodology (system testing and maintenance).

Phase seven (implementation and evaluating the system) entails the implementation and evaluation of the system, although only a concept demonstrator is developed in this study. For this reason the seventh phase is not documented in this thesis. There was, however, a formal discussion and exchange of information between the author and client so as to ensure that they are satisfied with the system and that the information provided by the author will be adequate to allow an external company to create and integrate a full DSS, based on the concept demonstrator presented in this thesis, into their systems should they desire it.

More detail on how each of the above seven phases of the SDLC were implemented is discussed in the sections that follow.

Note that for more than one of the above-mentioned phases, communication with the client and final user is emphasised. In order to demonstrate the extent to which the communication lines between the author and final user were open, a registry of all communication may be found in Appendix A.

6.2 Study objectives and user requirements

During the application of phase one (identifying problems, opportunities, and objectives) of the SDLC, the background of the study was assessed so as to identify needs and opportunities for improvement (and the removal of potential problems identified), and appropriate objectives were drawn up to exploit these opportunities. A discussion on the identified problems and list of objectives may be seen in Chapter 1.

During phase two (determining human information requirements) of the SDLC the aim was to gain an understanding of the needs of the final users of the system, thereby ensuring that identified opportunities for improvement are in line with the requirements of the final users. Although already discussed partially in Chapter 1, the two main user requirements obtained from a meeting between the systems analyst and final user are summarised as follows:

Investigate the variable importance of specific variables: Two different analyses are to be carried out in respect of the data collected for bursary applicants. First, the importance of variables currently used and considered as important by the main industry partner is investigated. Secondly, the importance of all other variables that have potential to be influential as success indicators is investigated.

Create a working DSS concept demonstrator: The main industry partner would like to see a working concept demonstrator of a DSS that may assist them in the process of allocating bursaries to potential candidates by considering whether or not they are expected to successfully graduate, based on pre-selected and configurable variables or factors. They will then be able to decide whether such a DSS would add adequate value to their organisation for them to further pursue the establishment of the full system.

6.3 Analysis of the system needs

Once the above-mentioned requirements and scope were well understood, diagrams were drawn up to depict the current state of the main industry partner's systems, referring specifically to its direct and indirect functions related to the applicants and the bursary application process. In order to do this, the UML's use case and activity diagrams (both of which were discussed in §2.3.6) were employed to visualise how the system should interact with its environment.

In order to meet the user requirements stated in §6.2, the focus in this section falls on the main industry partner's systems. The use case model and the activity diagram which follows in this section predominately depicts the industry partner's application process. Having said that, however, in both these diagrams, a system boundary is placed which delimits the scope of the DSS concept demonstrator developed. The area within the *system boundary*, as well as the DSS architecture previously presented in Chapter 5, is modular and may be reused within the context of other similar bursary providers.

6.3.1 Current bursary application process

Before proceeding with the presentation of the use case diagram and activity diagram, the current bursary application process of the main industry partner is elaborated upon. The selection process starts when a student completes a bursary application form, sometimes with the assistance of one of the main industry partner's recruiters (or regional representatives). The completed form, along with all other required documents and information, is sent on to the recruitment and selection department where an employee assigns a form score to the student. If the form score is high enough, then a recruiter returns to the student and conducts an interview with him or her. The recruiter assigns a interview score and recommendation to the student. All of the above information is then passed back to the main industry partner's selection panel which evaluates it and makes a final decision as to whether or not to award a bursary to the student.

6.3.2 Use case diagram

As mentioned in §2.3.6, a use case diagram enables the systems analyst to visualise the interactions between all the elements and activities of the desired system, thus obtaining a better overall view of the working of the system — with a focus on the connections between all the different role players. The role players for this use case are the *student* applying for the funding, the *main industry partner recruiter* who communicates directly with the student applying, the *main industry partner office secretary*, and the *main industry partner application committee* which evaluates the student's application. The use case derived in respect of the bursary application process discussed above, may be found in Figure 6.2. The dark grey block with a dashed outline indicates the system boundary within which the DSS demonstrator is designed.

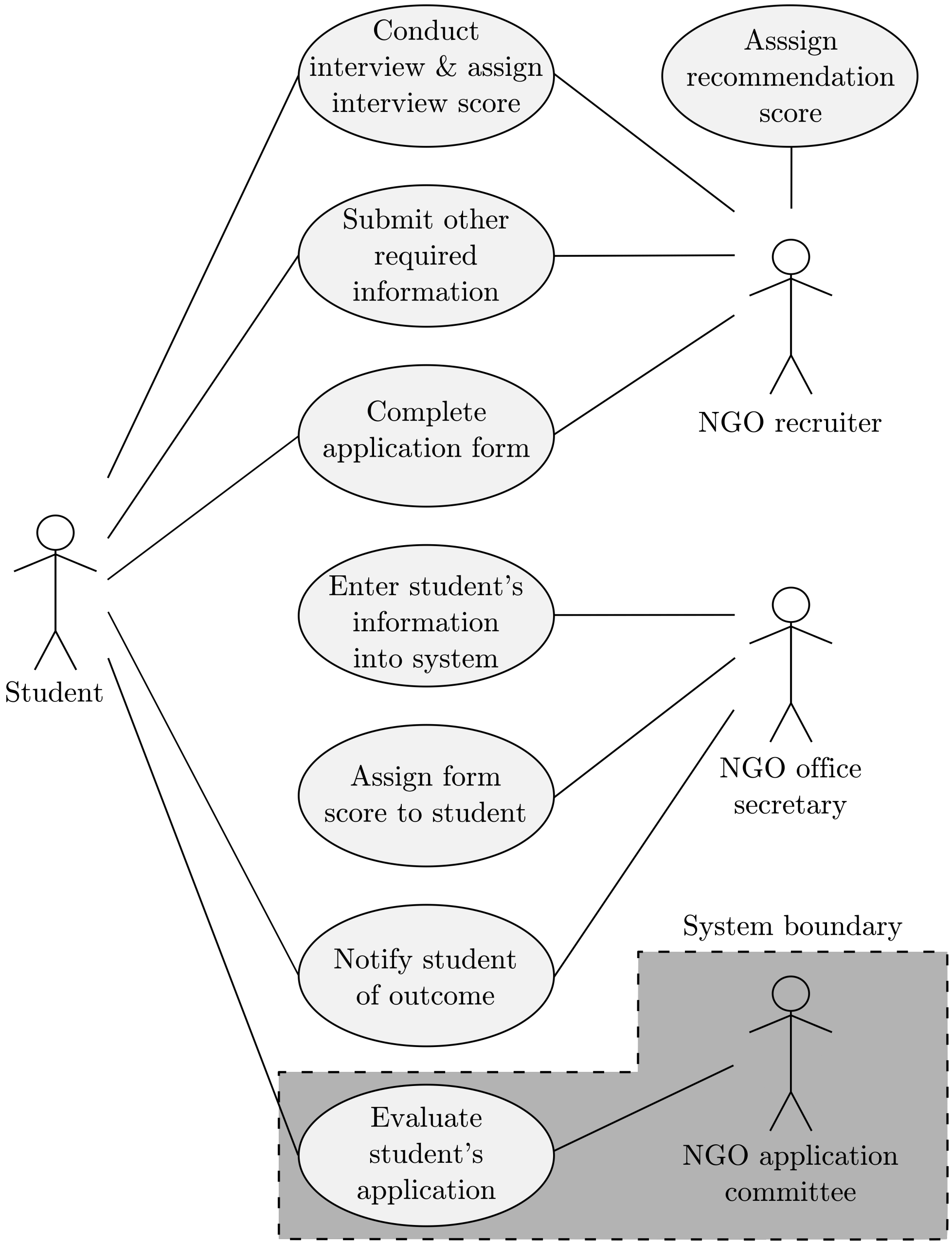

Conduct interview & assign interview score
Asssign recommendation score
Submit other required information
Complete application form
NGO recruiter
Student
Enter student's information into system
Assign form score to student
NGO office secretary
Notify student of outcome
System boundary
Evaluate student's application
NGO application committee

FIGURE 6.2: *Use case model for the main industry partner.*

6.3.3 Activity diagram

A good understanding of how the overall system operates may be obtained by visualising its parallel, branched and concurrent activity flows with the help of an activity diagram. The activity diagram for the system in question is shown in Figure 6.3. As before, the dashed outline indicates the system boundary within which the DSS concept demonstrator is designed.

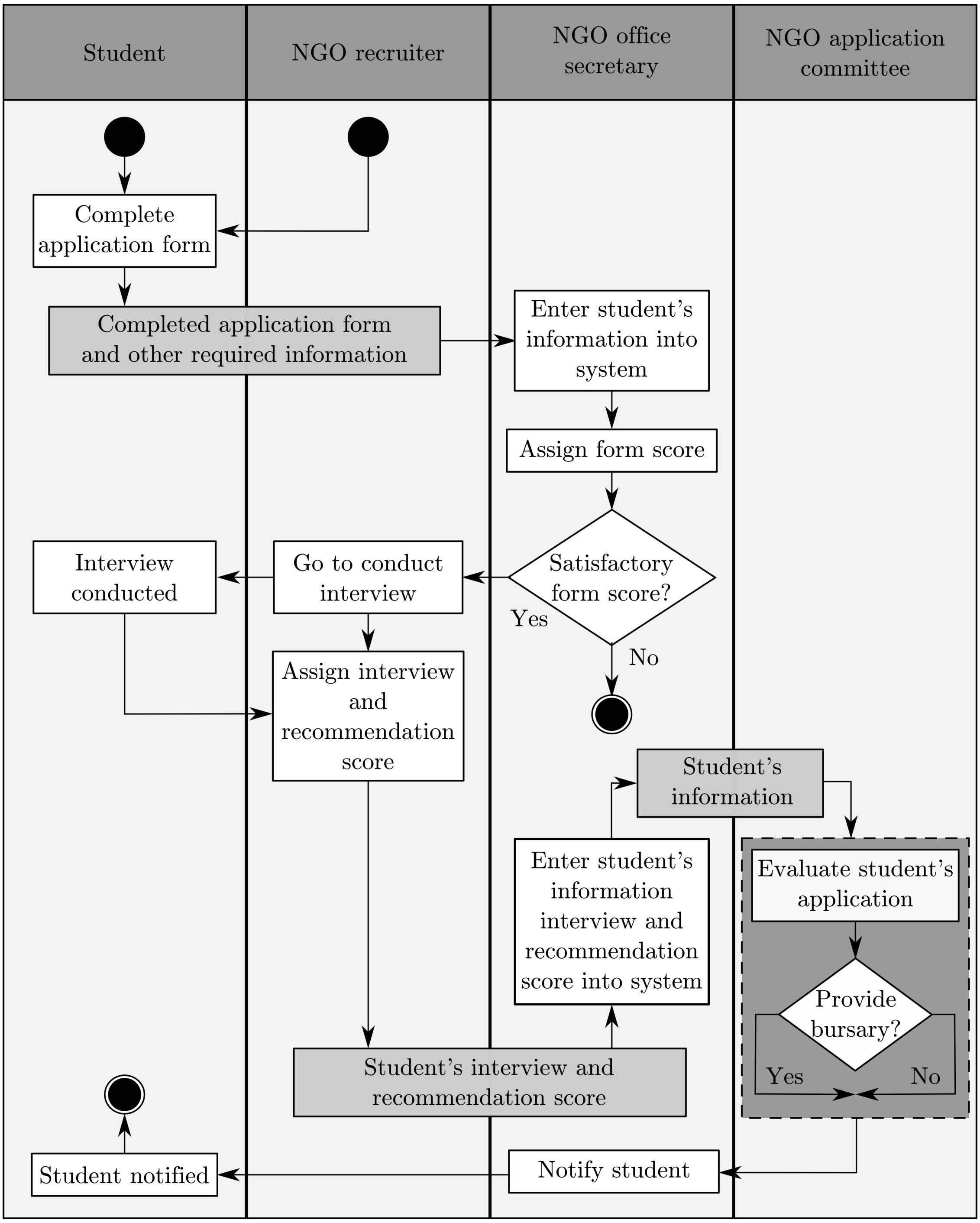

Student
NGO recruiter
NGO office secretary
NGO application committee
Complete application form
Completed application form and other required information
Enter student's information into system
Assign form score
Satisfactory form score?
Yes
No
Interview conducted
Go to conduct interview
Assign interview and recommendation score
Student's information
Enter student's information interview and recommendation score into system
Evaluate student's application
Provide bursary?
Yes
No
Student's interview and recommendation score
Student notified
Notify student

FIGURE 6.3: *Activity diagram of the process followed by the main industry partner.*

6.4 Design of the system

The fourth phase of the SDLC is concerned with the design of the system, focussing specifically on the establishment of an incarnation of the architecture of the DSS in question. The use case (presented in §6.3.2) and the activity diagram (presented in §6.3.3) illustrate how the system should interact with its environment, while the proposed architecture framework elucidates the inner workings of the system within the context of the system boundary. The framework architecture was designed and discussed in Chapter 5.

As the concept demonstrator of the DSS of Chapter 5 focuses on the model base, the creation of an additional database was deemed unnecessary. Rather, the only database within the architecture is that of the industry partner in question. That database provides `.csv` formatted `MS Excel` files, and the model base outputs files of the same format which may be used by the industry partner to pull the data back into its database. The internal structure of the industry partner's database was not disclosed to the author. It was agreed that the industry partner would only make use of the model base part of the DSS and not redesign their entire database based on the design of Chapter 5. In this manner, the DSS framework proposed in Chapter 5 becomes even more modular and is easily applicable to different NGOs.

6.5 Building of the system

The focus in this section is on how the DSS concept demonstrator was built. The section comprises five subsections. First, the data preparation steps are briefly highlighted, as well as how the statistical and data assumptions are validated. The three main framework elements of Chapter 5 implemented in the concept demonstrator are discussed thereafter. Next, the software, library, and human-computer interaction are considered. The final part of this section is devoted to a detailed walk-through of the DSS concept demonstrator so as to illustrate its capabilities.

6.5.1 Date preparation and assumptions

Recall that the notion of data preparation has previously been discussed in §2.2. This notion was also discussed in the context of the DSS in §5.2. The specifics of how the various data preparation steps are applied in the concept demonstrator is considered in the case study chapters that follow later in this thesis.

A tabular summary of various assumptions made in relation to the input and output data may be found in §3.11. Of those, the three assumptions of measurement error, mutual exclusivity of categorical variables, and the correct coding of the dependent variable have to be considered by the DSS concept demonstrator user before the data are pulled in by the DSS.

During data collection, the assumption of measurement error may be validated by following the guidelines presented in §3.2.1. The next two assumptions are validated during the data preparation step. For each potential pair of categories of a categorical variable, it should be ensured that the two categories are mutually exclusive. Next, for each potential dependent variable, the desired or positive outcome should be assigned the label 1, and the undesired outcome assigned the label 0. The result should be that all potential dependent variables in the data file imported into the DSS concept demonstrator contains only 0's and 1's, with the 1's associated with the desired outcome. Furthermore, all ordinal data fields, such as a rating system involving one star, two stars or three stars, should be converted to numeric values (*i.e.* 1, 2, and 3). The DSS concept demonstrator will thus be able to identify these fields as ordinal instead of nominal.

In addition to the past data file, a new data file of alternatives required to be ranked and have their outcome predicted should also be imported into the DSS concept demonstrator. The data preparation processes described above are also applicable to the new data and the new data file should also have the same column names as the past data file, with no new categories within the qualitative fields. For example, if within the data sets, there exists a field containing a list of tertiary institutions, there should be no new tertiary institutions listed within the new data file which are not also contained in the old data file. The DSS will be unable to predict an outcome for such students. Those specific cells (*i.e.* the ones containing the name of the new university) should rather be left blank, so that if that field (*i.e.* tertiary institution name) is selected as an independent variable, the entry of the student will be removed since there exists a blank. The student's record will, however, remain if fields other than the tertiary institution name field are selected as independent variables.

6.5.2 Selection of framework elements

The DSS framework requires a number of pre-specified learning base models, a static prediction weighting method, and a single MCDA outranking method to be selected for use within it, as described in §5.3. The selection of these elements are discussed in this section in the context of the concept demonstrator.

Note that other systems analysts may easily make use of the same proposed DSS framework presented in Chapter 5, but employ completely different framework elements. It is therefore emphasised that the elements chosen for implementation in this study are believed to be relatively good fits for the intended use of the DSS concept demonstrator by the main industry partner, but nevertheless represent subjective element choices. The aim is not to find the best elements, but rather to showcase the capability and flexibility of the DSS proposed in Chapter 5.

Selection of base model elements

The statistical learning base models selected as elements within the framework of Figure 5.4 are Logistic regression (described in §3.4), CART (described in §3.6), Random forests (described in §3.7), the C4.5 algorithm (described in §3.8), and SVMs (described in §3.9).

In order to further expand on the specific libraries and the functions of these libraries employed within `R` to implement these base models, which are provided in Appendix B, a summary of how each base model is applied now follows.

Logistic regression is performed according to a binomial error distribution. The probability outcome for each observation is converted to a binary outcome, taking 0.5 as the cut-off value. Those observations with a probability outcome of 0.5 or above receive the label 1, and those less than 0.5 receive the label 0.

CART is performed by first growing a maximum tree with the cost complexity set to zero, adopting the gini splitting criterion. Next, the best cost complexity parameter is determined for pruning a maximum tree back, using the method discussed in §3.6. This is repeated ten times for ten maximum trees using 10-fold cross-validation. A final maximum tree is grown based on all the learning data and is subsequently pruned back using the averaged cost complexity parameter determined from the previous step.

The method of random forests is used to build 15 000 trees (*i.e.* selecting the value $ntree = 15\,000$) adopting the gini splitting criterion and with $mtry = \sqrt{m}$ (rounded to the nearest integer), where m is the number of independent variables. The trees are built to the maximum size that the minimum node size (number of observations) allows by setting the minimum terminal node size to three.

The tree building algorithm C4.5 is employed to build a single decision tree according to the information gain ratio splitting criterion to the maximum size allowed by setting the minimum number of observations per terminal node to two. Thereafter, the technique of reduced error pruning is used to trim the decision tree.

The SVM kernel adopted is the Gaussian radial basis function, and its parameter, σ, is determined using the heuristic presented by Caputo *et al.* [36]. Finally, the error penalty is set to the default value of 1, as suggested by Karatzoglou *et al.* [101].

Selection of the integration strategy element

The different integration strategies which may be used to combine the predictions of different base models into a single prediction for each alternative were described in §4.2.1. In the concept demonstrator, an extension of *weighted majority voting* is applied. This involves not only assigning a higher weighting to those base models which perform better than the other base models in predicting the outcome of students (which is the process followed for weighted majority voting), but also slightly boosts the weights up or down so as to better pronounce the differentiation between the best and worst models. The weights of the base models which perform average in comparison to the other models all remain the same, while those of the models which perform better are slightly increased, or slightly decreased for those models which perform below average. The sum of these weights is still one, as required by the ELECTRE III method. The reason for this weighting adjustment is that after the normally weighted majority voting, the weights are reasonably close to one another, resulting in a situation where those base models which perform well are not adequately rewarded.

Note that any base model which produced a prediction accuracy of less than 57% on the validation set was excluded and thus not used in the calculation of the combined prediction or ranking of alternatives. Upon investigation of the effect of this cut-off percentage, it was found that a cut-off value of 55% or smaller did not have the desired positive effect and a cut-off value greater than 60% removed too many models for some samples — thus, the midpoint of 57% was chosen. This was necessary since for the ensemble method to function properly it should be comprised of base models which performed at least slightly above average.

Selection of the MCDA element

An approach toward selecting a single MCDA method or set of MCDA methods suitable for use in the context of a specific decision situation was described in §4.3.6. The four-stage typology tree of that approach is applied in this section to identify suitable MCDA methods for use in a case study involving data provided by the main industry partner. The four stages of the typology tree, along with the selections made during each stage, may be seen in Figure 6.4. A justification of the selections follows. As the industry partner is interested in ranking bursary applicants, the *Outranking synthesizing approach* is selected from the operational approach options during stage one, and *Ranking* is selected from the problematic options during stage four. The input information for the base models is quantitative in nature, as they produce binary predictions. The option *Quantitative* is therefore selected during stage two. Since the nature of the input information is deterministic, and the output should be reproducible and consistent, *Deterministic* is selected during stage three.

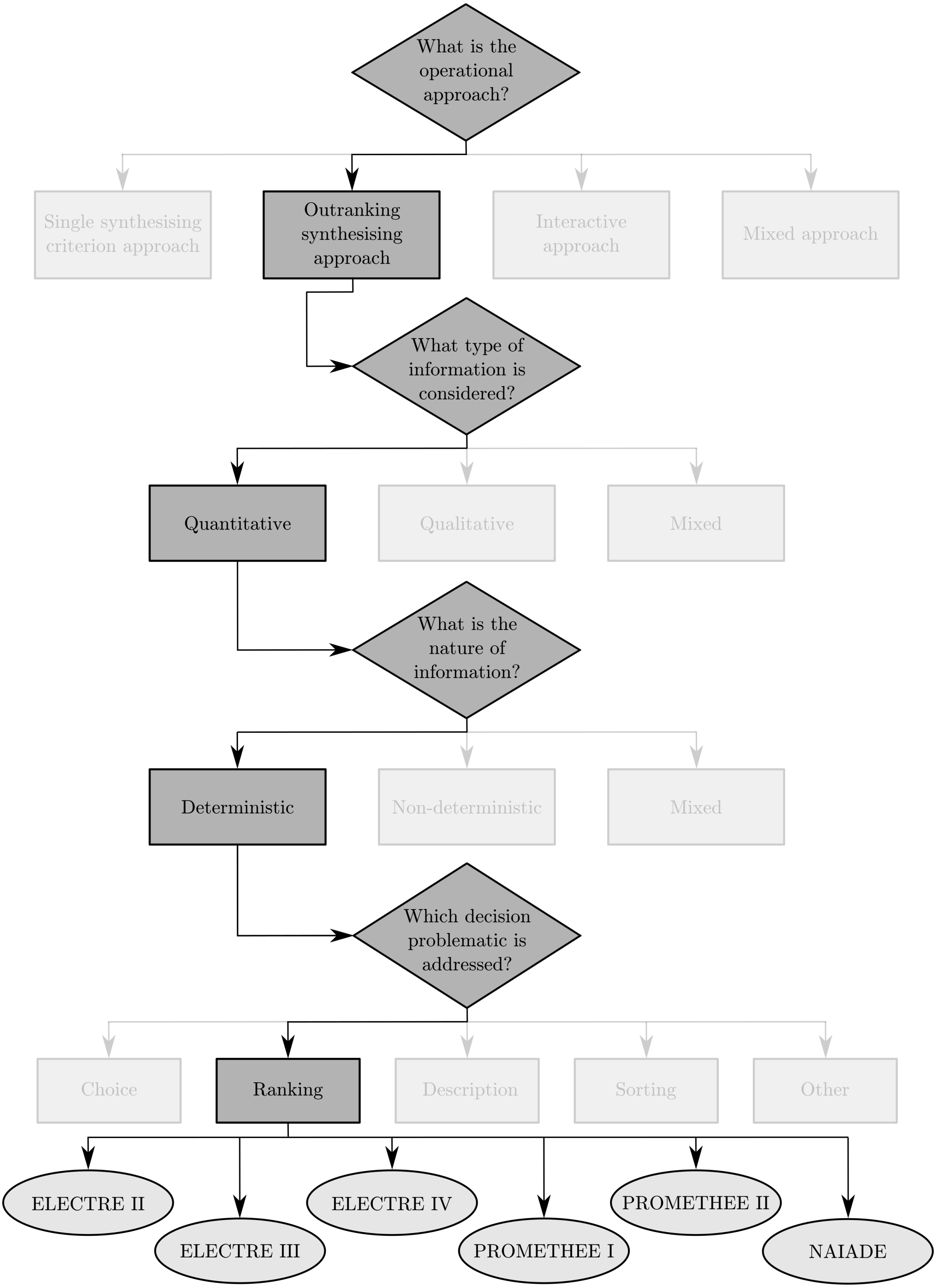

What is the operational approach?
Single synthesising criterion approach
Outranking synthesising approach
Interactive approach
Mixed approach
What type of information is considered?
Quantitative
Qualitative
Mixed
What is the nature of information?
Deterministic
Non-deterministic
Mixed
Which decision problematic is addressed?
Choice
Ranking
Description
Sorting
Other
ELECTRE II
ELECTRE III
ELECTRE IV
PROMETHEE I
PROMETHEE II
NAIADE

FIGURE 6.4: *The four stages of Mota's typology tree, as applied in the context of the current study.*

From the original list of twenty eight MCDA methods presented in §4.3.4, only six remain after having carried out the selection process described above. The six remaining methods are listed in Table 6.1. These methods may, however, be refined further, as follows. According to Figueira *et al.* [66, p. 167], ELECTRE III was designed as an improvement of ELECTRE II. They also state that the popular ELECTRE III has been applied with marked success over the past two decades, achieving its purpose. For this simple reason, ELECTRE II is not considered for implementation in the case study. From the ELECTRE family, ELECTRE III and ELECTRE IV therefore remain. Only ELECTRE III is, however, utilised because, as Tzeng and Huang [233, p. 286] point out, ELECTRE IV is a similar and simplified version of ELECTRE III. In addition, Ishizaka and Nemery [97] and Tam *et al.* [222, p. 47] are in agreement that ELECTRE III is the most popular ELECTRE method and is thus the logical choice between the two.

TABLE 6.1: *MCDA methods suitable for use in the current study.*

MCDA method	Guideline 3: Problematic	Guideline 4.1: Kind of information			Guideline 4.2: Information features	
		Qualitative	*Quantitative*	Mixed	*Deter.*	Non-deter.
Outranking methods						
ELECTRE II	γ	✓	✓	✓	✓	
ELECTRE III	γ	✓	✓	✓	✓	
ELECTRE IV	γ	✓	✓	✓	✓	
PROMETHEE I	γ	✓	✓	✓	✓	
PROMETHEE II	γ	✓	✓	✓	✓	
NAIADE	γ	✓	✓	✓	✓	✓

The NAIADE method is also excluded for the simple reason that it does not utilise weights assigned to the criteria by the decision maker, as opposed to the ELECTRE and PROMETHEE methods which do accommodate the assignment of weights [50, p. 107]. Brans and de Smet [28, p. 168] point out that assigning weights to a criterion is not a simple process — this process is associated with many complications, including bias of the decision maker. The framework to be employed, as discussed in Chapter 5, however, utilises weights determined according to a well-structured approach as opposed to adopting objective estimations by managers. For this reason, methods which directly utilise weights are preferred in the context of this study, so as not to waste valuable input data.

The result of the above exclusions, only the MCDA methods ELECTRE III, PROMETHEE I, and PROMETHEE II remain. As was mentioned at the start of this section, selection of the elements for the framework of the current study were not made on the basis of using the single best elements, but rather to showcase the capabilities and flexibility of the framework. The same logic applies here. In the case of the base models, more than one element may be added to the framework, but for the ranking method only one element may be employed, and thus only one of the three candidate outranking methods may be selected. Although any of these methods may be used as the MCDA method in the framework of Chapter 5, the logical choice is the best known and documented method of the remaining three candidates, ELECTRE III, which was described in detail in §4.3.5.

As pointed out in §4.3.5, there are often no specific guidelines for selecting ELECTRE III thresholds for specific criteria of a certain field (*e.g.* environmental science), in which case three general ground rules may be applied. First, the relationship between the ELECTRE III thresholds $q \leq p \leq v$ should be enforced. Next, p should be substantially larger than q, which may be as much as double or even more. Lastly, v may take the same value as p, but it can be up to three to five times larger than p. In addition, the difference between v and p should decrease as the importance weighting of a specific criterion increases. By use of the proposed DSS, the range of all criteria will be equal (*i.e.* between zero and one). The result is that the threshold of each criterion will fall within the same range of between zero and one. Using these three ground rules, the following relationship between thresholds is selected for use in ELECTRE III. First, q was chosen as 0.1 and p was chosen as 0.2 (twice the value of q). The value of v is varied between 0.5 and 0.9, based on the weight of each criterion (the higher the weight, the lower v will be), thus ensuring that the value of v varies between being 2.5 and 4.5 times larger than p.

6.5.3 Software and library considerations

In §2.3.2 it was mentioned that the six categories of *license*, *connectivity*, *data store*, *language on server side*, *integrated development environment*, and *web application framework* are some of the more important ones that have to be considered when selecting a specific software suite for DSS concept demonstrator development purposes. For the current study, the selections made for each of the six categories may be seen in Figure 6.5.

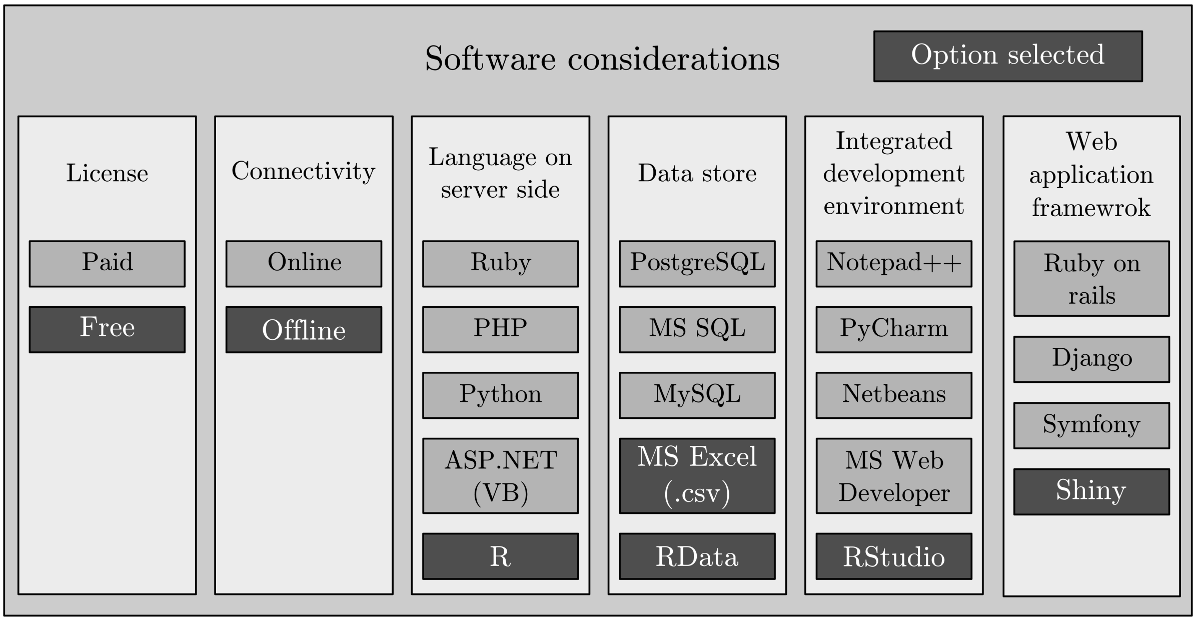


FIGURE 6.5: *Software considerations in the development of the DSS concept demonstrator.*

The reasons for the selections in Figure 6.5 are as follows. Due to financial constraints on the side of the main industry partner, it was decided that software with a *free* licence should be employed. As the DSS will function as a concept demonstrator, it is not necessary to have it operating online and thus its connectivity will be *offline*. Although more than one server side language could have been utilised to successfully complete the study, the selection of `R` was based on the following main reasons. Besides the fact that `R` is free, there exists a large and knowledgeable online community, including various colleagues of the author, which utilises `R` and hence the necessary support structure is available when selecting `R`. Also, `R` is statistically orientated and many good pre-built packages exists for use within it. The data store type `R Data` is directly linked with `R` and thus compulsory. The second data store choice of `MS Excel's .csv` file type was justified in §6.4. Next, the choice of `R Studio` as the IDE was the obvious choice as it is the most popular and a well-supported IDE associated with `R`. Similarly, `Shiny` is the most popular web application framework and thus also the obvious choice. The specific libraries and functions of the libraries that are employed in `R` in this study are specified in Appendix B.

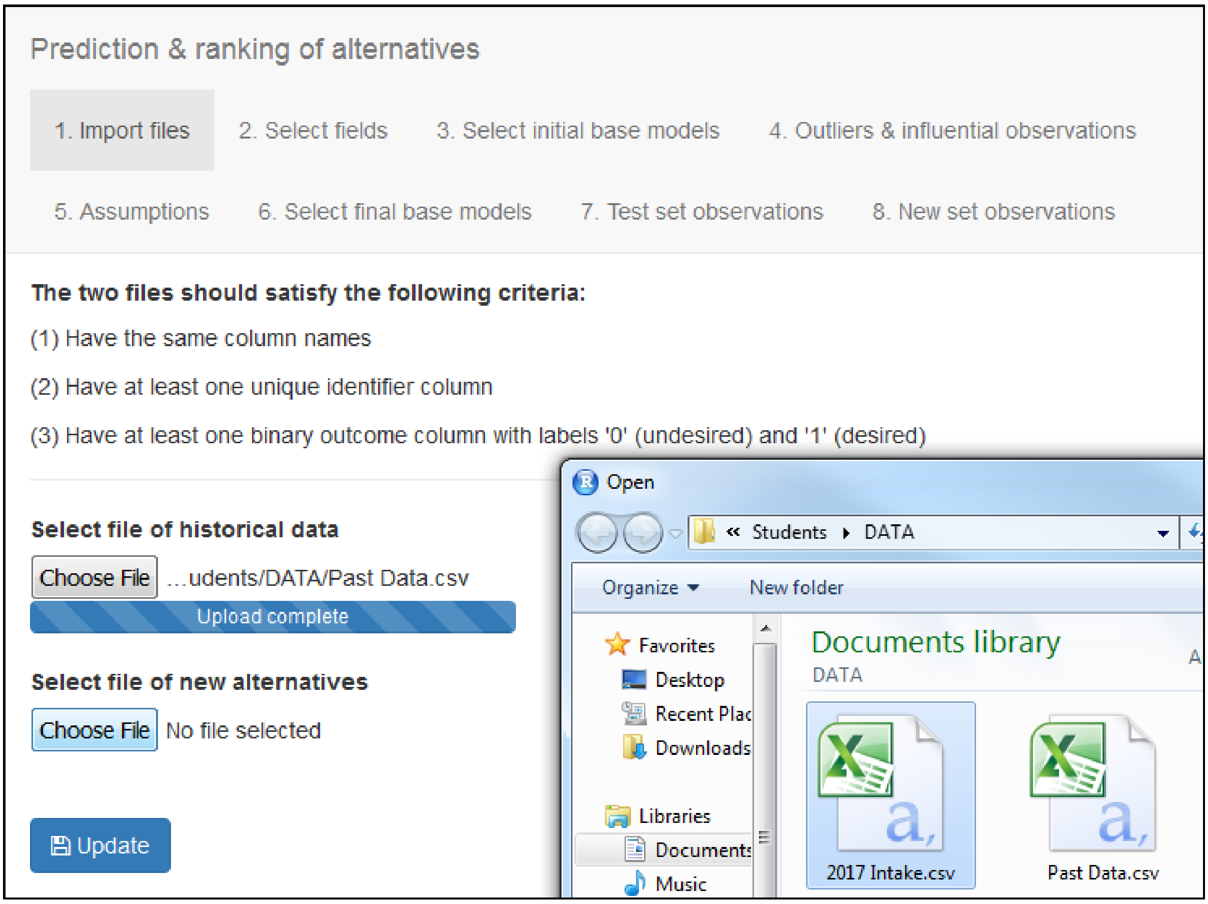
6.5.4 Human-computer interaction considerations

The topic of human-computer interaction was discussed in §2.3.2. It was mentioned that usability is the most important aspect of a system's interface. Mandel [133] also identified three golden rules that should be abided by during the HCI development process when creating a usable interface. They are: place the user in control, make the user interface consistent and reduce the user's memory load. These were of primary concern to the author when designing the system described here. As mentioned, the iterative user interface design and evaluation process of Figure 2.3 was used in this study to involve the final user and obtain valuable feedback. Adoption of this iterative process during the design of the DSS and GUI is evident from the communication registry in Appendix A. The description in the next section makes use of screen shots to present a visual walk-through explanation of the DSS created using `Shiny`. From this walk-through discussion it should become clear that the author has made an effort to apply the above-mentioned three rules, which will again be discussed after the walk-through.

6.5.5 System walk-through

Using actual data provided by the main industry partner, the working and capabilities of the proposed DSS are demonstrated in this section in a walk-through manner. The process to be followed by the user has been partitioned into eight main steps. The screen associated with the first step may be seen in Figure 6.6.

Each step is associated with one of the eight numbered tabs at the top of the GUI screen. The grey block containing the eight tabs is permanently visible at the top of the screen on all eight



The two files should satisfy the following criteria:

(1) Have the same column names

(2) Have at least one unique identifier column

(3) Have at least one binary outcome column with labels '0' (undesired) and '1' (desired)

Select file of historical data

Choose File ...udents/DATA/Past Data.csv

Upload complete

Select file of new alternatives

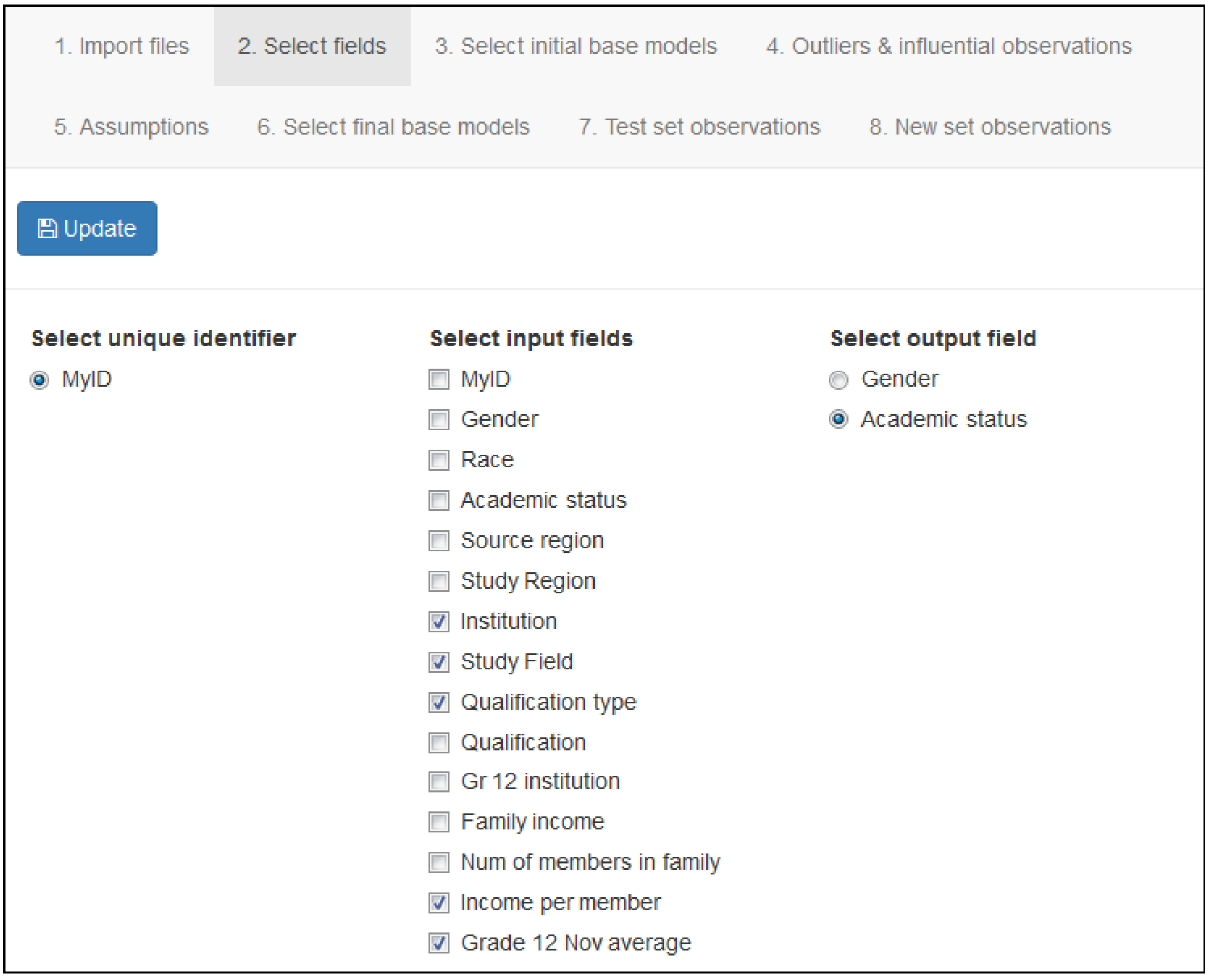
Choose File No file selected

FIGURE 6.6: *Tab one: Files of input data are selected for importation into the system.*

screens produced by selecting any of the tabs. The user may sequentially proceed to each subsequent tab once they have completed the steps associated with the current one. At any stage, however, the user is able to return to a previous tab by selecting it.

On the GUI screen of the first tab, the user is prompted to select two files of data for importation. The first should contain past data from which the system will learn, and the second should contain data of new alternatives for which the output or predictions are not known, but desired. The two files should satisfy the three conditions listed on the screen, namely that the files should contain the same column names, have at least one column containing unique identifiers, and have at least one binary outcome column with labels “0” and “1.” The GUI informs the user once the two files have successfully been uploaded, after which the user may select the “Update” option before proceeding to tab two.

On the screen associated with tab two, shown in Figure 6.7, the user is required to provide three inputs. First, the user has to select exactly one field to be used as the unique identifier of alternatives for the remainder of the analysis. Secondly, the user may select multiple fields to be taken as the independent variables or input variables for the analysis. Finally, the user may select a single binary field to be used as the dependent variable or output variable. In order to assist the user, the system only provides those fields that are viable options for each of the three specific inputs.



Select unique identifier

● MyID

Select input fields

□ MyID
□ Gender
□ Race
□ Academic status
□ Source region
□ Study Region
☑ Institution
☑ Study Field
☑ Qualification type
□ Qualification
□ Gr 12 institution
□ Family income
□ Num of members in family
☑ Income per member
☑ Grade 12 Nov average

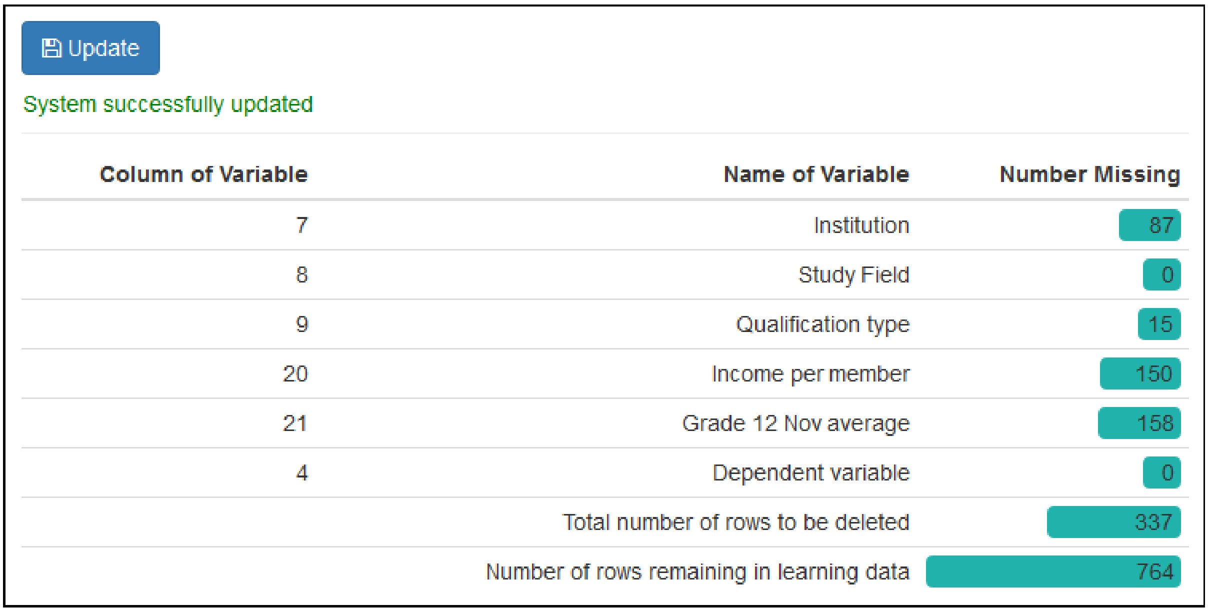
Select output field

○ Gender
● Academic status

FIGURE 6.7: *Tab two: Selection of a unique identifier, input variables, and an output variable from the available fields of the imported data.*

In Figure 6.7, the *MyID* column has been selected as the unique identifier of the students and the *Academic status* of a student has been selected to be the binary output variable. Five independent variables have furthermore been selected, namely *Institution*, *Study field*, *Qualification type*, *Grade 12 November average*, and *Family income per member*.

Once the user is satisfied with his or her selection, the "Update" button may be clicked so as to update the system. The system will then produce on the same tab, a table containing the number of entries, observations or alternatives to be deleted due to missing values in any of the fields or columns selected, as shown in Figure 6.8. The total number of observations deleted from the training data thus depends on the selection of variables. The user is therefore able to experiment with the variable selection so as to ascertain its effects in respect of the remaining data. Once satisfied with his or her selection of variables and the number of observations remaining after deletion of missing values, the user may proceed to the third tab.



System successfully updated

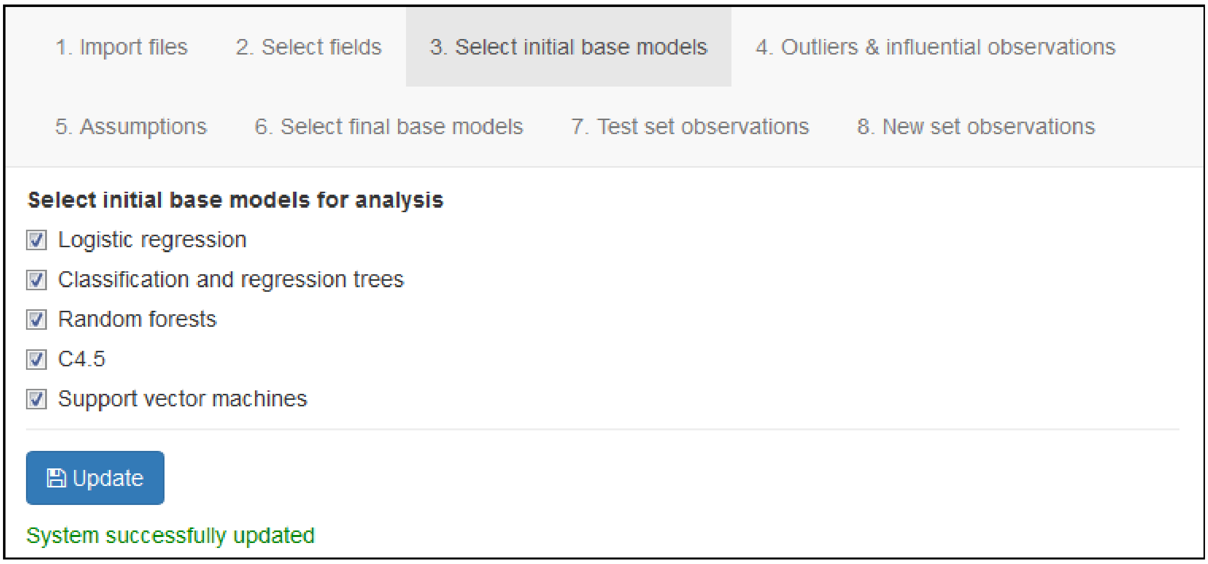
Column of Variable	Name of Variable	Number Missing
7	Institution	87
8	Study Field	0
9	Qualification type	15
20	Income per member	150
21	Grade 12 Nov average	158
4	Dependent variable	0
	Total number of rows to be deleted	337
	Number of rows remaining in learning data	764

FIGURE 6.8: *Tab two: Number of data entries to be deleted due to missing values.*

In the third tab, shown in Figure 6.9, the base models currently available in the DSS are listed and the user may select those that he or she would like to use for analysis purposes. In Figure 6.9, all of the available base models have been selected. As before, the user may click the "Update" button so as to update the system before proceeding the the next tab.

In the fourth tab, shown in Figure 6.10, the user may identify the outliers and influential observations within the data by clicking the "Locate" button, at which point the system will produce a table indicating the number of outliers and influential observations. Once the user has assessed the results thus tabulated, he or she may choose whether or not to delete the outliers and influential values. Suppose the user feels that too large a proportion of the observations would be removed if the outliers were to be deleted and hence does not elect to delete the outliers.

Note that when the "Locate" button is initially pressed the system partitions the sample of historical data into a learning set and validation set, as was shown in Figure 3.12. Only the learning set is scanned for outliers and influential values, as doing so for the validation set as well would increase the prediction assessment of any model in respect of such a test set cleaned of outliers significantly, potentially producing misleading predictive accuracy results. The partitioning of the sample data into a learning set and a validation set is achieved by

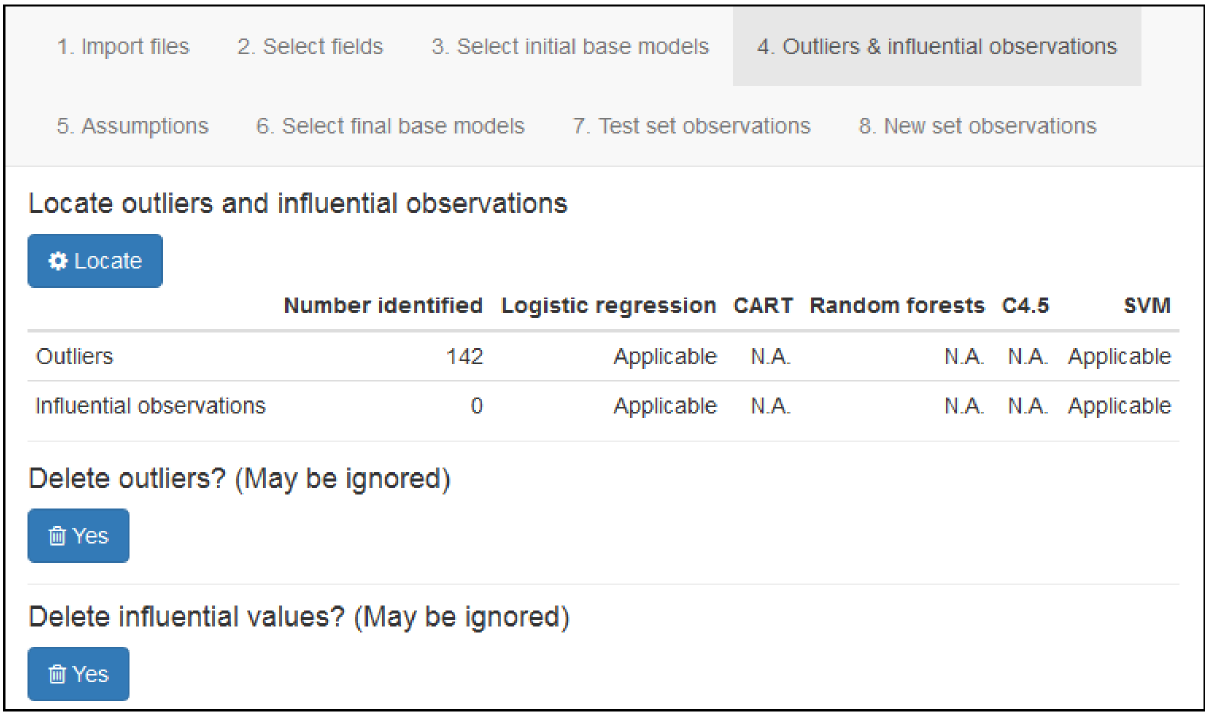


Select initial base models for analysis

☑ Logistic regression
☑ Classification and regression trees
☑ Random forests
☑ C4.5
☑ Support vector machines

System successfully updated

FIGURE 6.9: *Tab three: Initial selection of base models available to the user.*



Locate outliers and influential observations

	Number identified	Logistic regression	CART	Random forests	C4.5	SVM
Outliers	142	Applicable	N.A.	N.A.	N.A.	Applicable
Influential observations	0	Applicable	N.A.	N.A.	N.A.	Applicable

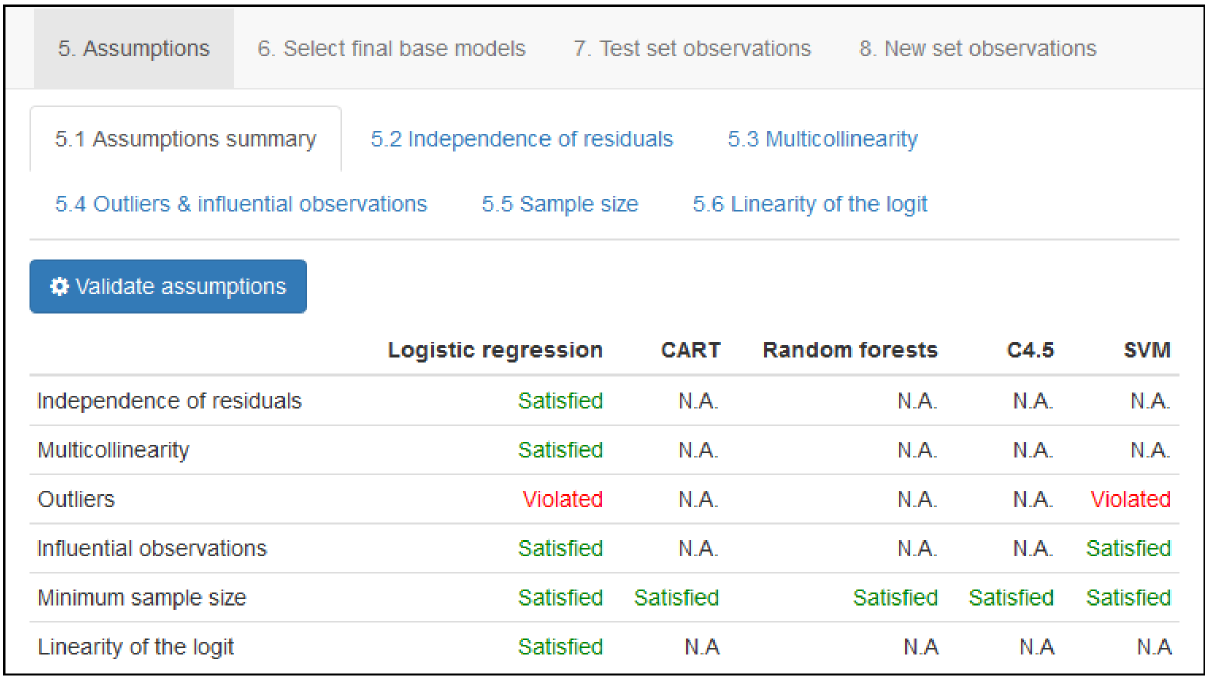
Delete outliers? (May be ignored)

Delete influential values? (May be ignored)

FIGURE 6.10: *Tab four: Identification of outliers and influential observations.*

randomly selecting observations based on a randomised seed value. The validation set size also varies randomly between 15% and 20% of the sample data. Each of the base models will further partition the learning set into training and test sets as per their specifications.

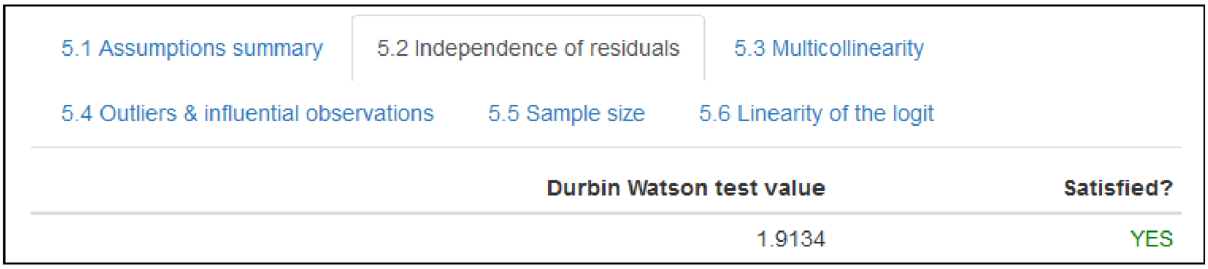
Once the user has completed the tasks in tab four (Figure 6.10), he or she may proceed to the fifth tab, shown in Figure 6.11, in which the user is presented six sub-tabs. Recall that in tab four the assumption was made that outliers and influential values would not be deleted. The statistical and data assumptions applicable to the various base models may be assessed and validated in the "Assumptions summary" sub-tab by clicking the "Validate assumptions" button. This will produce a table listing all assumptions applicable to each of the base models and whether these assumptions are satisfied or violated.



	Logistic regression	CART	Random forests	C4.5	SVM
Independence of residuals	Satisfied	N.A.	N.A.	N.A.	N.A.
Multicollinearity	Satisfied	N.A.	N.A.	N.A.	N.A.
Outliers	Violated	N.A.	N.A.	N.A.	Violated
Influential observations	Satisfied	N.A.	N.A.	N.A.	Satisfied
Minimum sample size	Satisfied	Satisfied	Satisfied	Satisfied	Satisfied
Linearity of the logit	Satisfied	N.A	N.A	N.A	N.A

FIGURE 6.11: *Tab five: Assumptions — Overview table.*

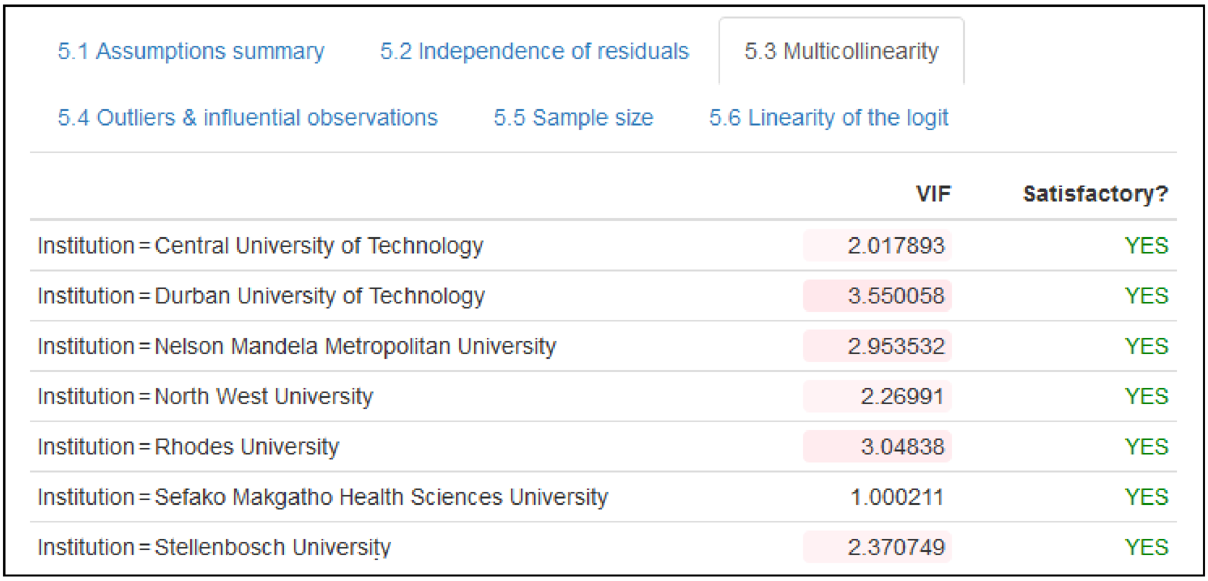
They user may investigate the assumption results in more detail, if desired, by clicking each of the remaining five sub-tabs, upon which the exact values determined for each assumption are shown, as depicted in Figures 6.12–6.16. More specifically, the Durbin-Watson coefficient value for the independence of residuals, the VIF scores for each of the independent variables (each category of a qualitative variable becomes a variable for this assessment), the number of outliers and influential observations identified (the same as was shown in tab four), the minimum sample size requirements of the different base models, and the linearity of the logit outcome for each quantitative and ordinal variable are displayed.


5.1 Assumptions summary
5.2 Independence of residuals
5.3 Multicollinearity
5.4 Outliers & influential observations
5.5 Sample size
5.6 Linearity of the logit

Durbin Watson test value	Satisfied?
1.9134	YES

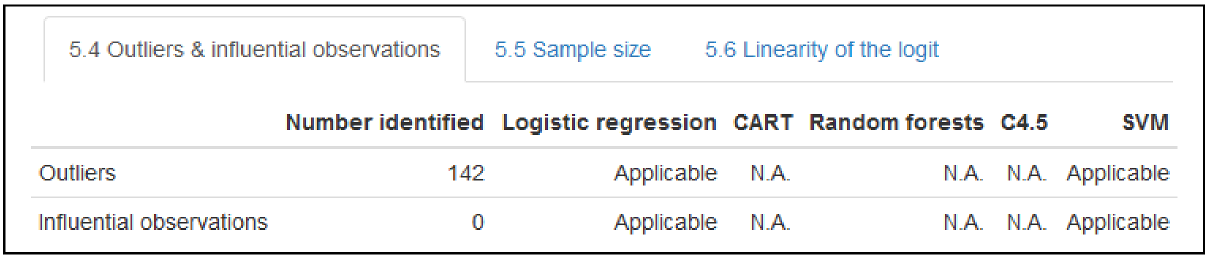
FIGURE 6.12: *Tab five: Assumptions — Independence of residuals.*

Once the user has assessed the assumption details, he or she may proceed to the next main tab, tab six, as shown in Figure 6.17. Armed with the knowledge of the assumption assessment results, the user may then make his or her final selections as to which of the base models should be included in the analysis. In this manner the user is placed in control, but is also made aware of the potential risks associated with his or her choices. Suppose, for illustration purposes, that the user is comfortable with the risk involved in retaining the outliers and still selecting both the Logistic regression and Support vector machine base models along with the other three base models, as shown in Figure 6.17. Once the user is satisfied with his or her choices, the "Update" button may be clicked to update the system before proceeding to the next tab.


5.1 Assumptions summary
5.2 Independence of residuals
5.3 Multicollinearity
5.4 Outliers & influential observations
5.5 Sample size
5.6 Linearity of the logit

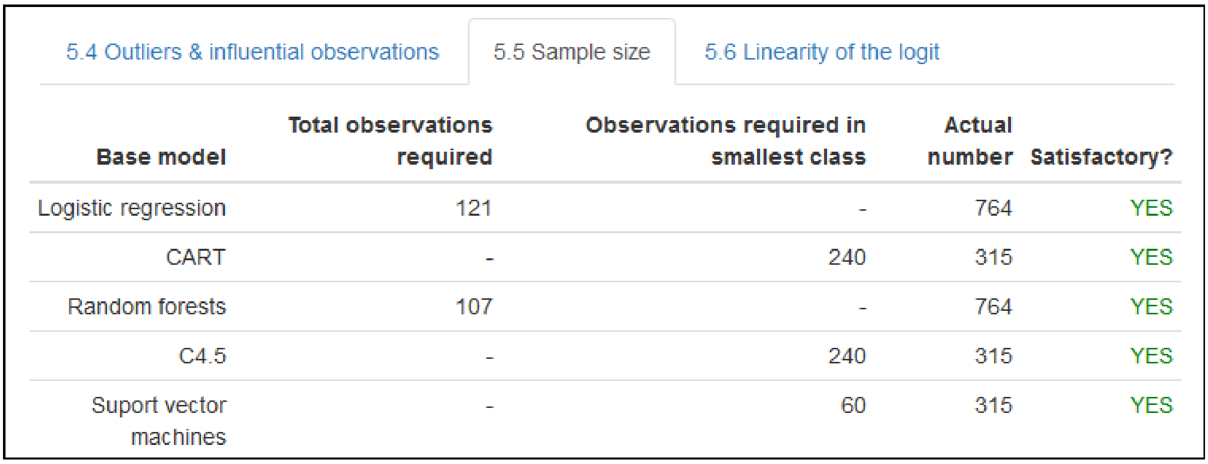
	VIF	Satisfactory?
Institution = Central University of Technology	2.017893	YES
Institution = Durban University of Technology	3.550058	YES
Institution = Nelson Mandela Metropolitan University	2.953532	YES
Institution = North West University	2.26991	YES
Institution = Rhodes University	3.04838	YES
Institution = Sefako Makgatho Health Sciences University	1.000211	YES
Institution = Stellenbosch University	2.370749	YES

FIGURE 6.13: *Tab five: Assumptions — Multicollinearity.*


5.4 Outliers & influential observations
5.5 Sample size
5.6 Linearity of the logit

	Number identified	Logistic regression	CART	Random forests	C4.5	SVM
Outliers	142	Applicable	N.A.	N.A.	N.A.	Applicable
Influential observations	0	Applicable	N.A.	N.A.	N.A.	Applicable

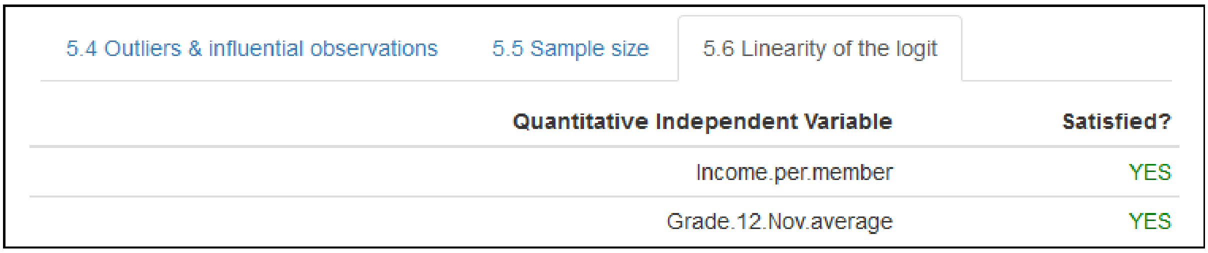
FIGURE 6.14: *Tab five: Assumptions — Outliers and influential values.*


5.4 Outliers & influential observations
5.5 Sample size
5.6 Linearity of the logit

Base model	Total observations required	Observations required in smallest class	Actual number	Satisfactory?
Logistic regression	121	-	764	YES
CART	-	240	315	YES
Random forests	107	-	764	YES
C4.5	-	240	315	YES
Suport vector machines	-	60	315	YES

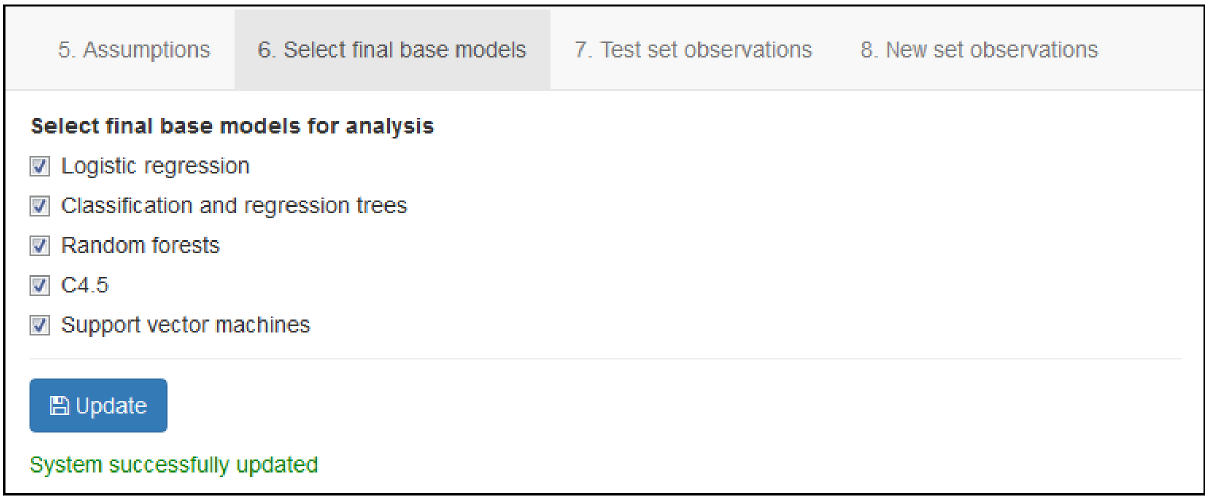
FIGURE 6.15: *Tab five: Assumptions — Minimum sample size.*

The system is now ready to predict the outcome for and rank each of the alternatives in the validation set of the sample of historical data. To achieve this, the system performs two operations. First, based on the predictive accuracy of each base model (measured by training the models in the context of the learning data and testing the results in the context of the unseen validation data), each model is assigned a weight, which is used to produce the single weighted vote for each alternative or student. In addition, an MCDA outranking method, in this case ELECTRE III, is used to assign each alternative to a ranking level.


5.4 Outliers & influential observations
5.5 Sample size
5.6 Linearity of the logit

Quantitative Independent Variable	Satisfied?
Income.per.member	YES
Grade.12.Nov.average	YES

FIGURE 6.16: *Tab five: Assumptions — Linearity of the logit.*



Select final base models for analysis

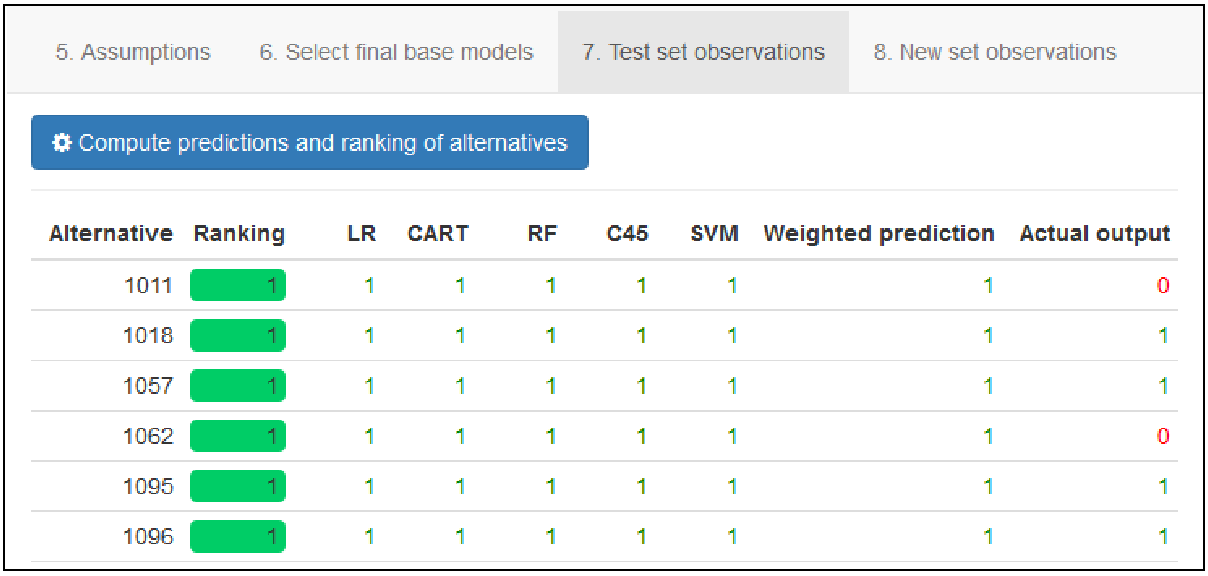
☑ Logistic regression
☑ Classification and regression trees
☑ Random forests
☑ C4.5
☑ Support vector machines

System successfully updated

FIGURE 6.17: *Tab six: Selecting the final base models.*

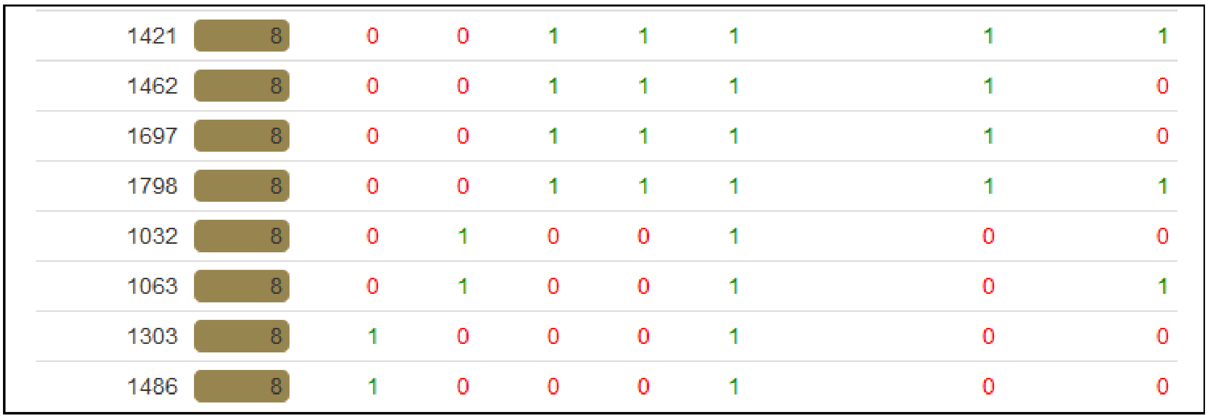
Suppose twelve rank levels are identified for some set of sample data. A portion of the results of the process described above may be seen in Figures 6.18–6.20. The weighted prediction for each alternative is shown in the second last column and the rank level of each alternative in the second column. Notice that the highest rank level is indicated on a light green background and that the lowest rank level is indicated on a red background, while all those levels in between are shaded between these two colours. The prediction accuracy of and weight for each of the base models study may be seen in Figure 6.20. Note that the predictive accuracy of the weighted prediction, of 66.2%, is higher than any of the single base models' predictive accuracies. For the majority of cases there will exist a single partition of the 1's and 0's of the weighted predictions column. The term 'partition' in this instance refers to the point at which, as one moves down from the most likely to succeed to the least likely to succeed applicants, at which the weighted prediction of the applicants changes from 1's to 0's. In Figure 6.19, for example the partition occurs within rank eight. The user may yet again proceed to the next tab, tab eight, once he or she has studied the information in tab seven.

The system is now ready to perform its final operation, namely predicting the outcome for and ranking the new alternatives (the second filed uploaded in tab one) for whom the output value are not known. This is the first time the new data set is used. The selected base models, which have been trained according to the learning data, now make predictions for the new data. The predictions from the various base models are combined using the weightings of the base models determined during the learning process. Finally, each new alternative is also assigned a rank level using the MCDA ELECTRE III outranking method. This final prediction and ranking of the new alternatives (this time on thirteen levels) may be seen in Figures 6.21–6.23. If the user is satisfied with the results, he or she may click the "Export output to Excel" button, shown in Figure 6.21, to save the results of all eight tabs to a single `MS Excel` workbook.


5. Assumptions
6. Select final base models
7. Test set observations
8. New set observations
Compute predictions and ranking of alternatives

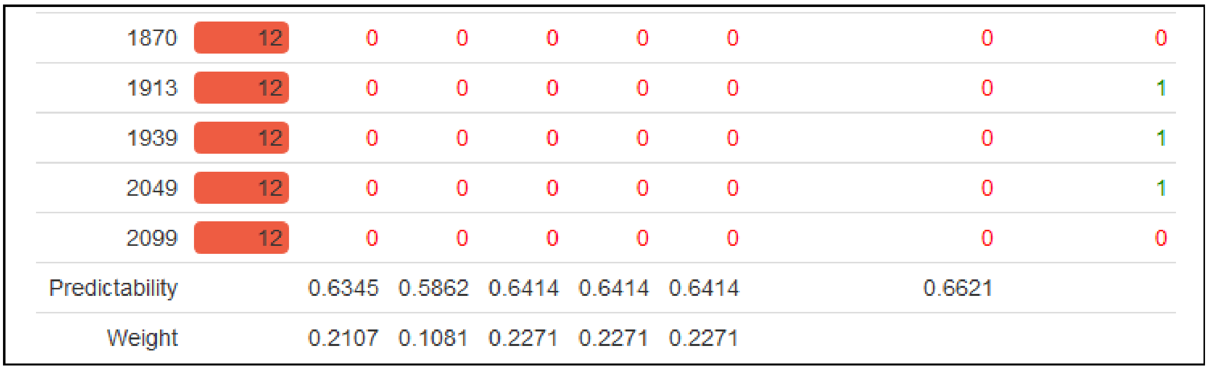
Alternative	Ranking	LR	CART	RF	C45	SVM	Weighted prediction	Actual output
1011	1	1	1	1	1	1	1	0
1018	1	1	1	1	1	1	1	1
1057	1	1	1	1	1	1	1	1
1062	1	1	1	1	1	1	1	0
1095	1	1	1	1	1	1	1	1
1096	1	1	1	1	1	1	1	1

FIGURE 6.18: *Tab seven: Prediction and ranking of test set data (top).*



1421	8	0	0	1	1	1	1	1
1462	8	0	0	1	1	1	1	0
1697	8	0	0	1	1	1	1	0
1798	8	0	0	1	1	1	1	1
1032	8	0	1	0	0	1	0	0
1063	8	0	1	0	0	1	0	1
1303	8	1	0	0	0	1	0	0
1486	8	1	0	0	0	1	0	0

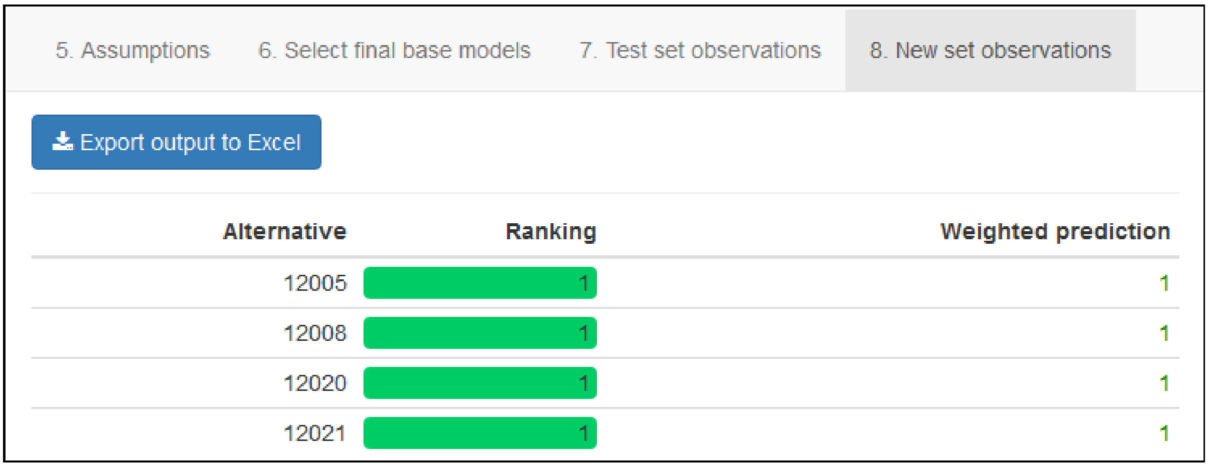
FIGURE 6.19: *Tab seven: Prediction and ranking of test set data (middle).*



1870	12	0	0	0	0	0	0	0
1913	12	0	0	0	0	0	0	1
1939	12	0	0	0	0	0	0	1
2049	12	0	0	0	0	0	0	1
2099	12	0	0	0	0	0	0	0
Predictability		0.6345	0.5862	0.6414	0.6414	0.6414	0.6621	
Weight		0.2107	0.1081	0.2271	0.2271	0.2271		

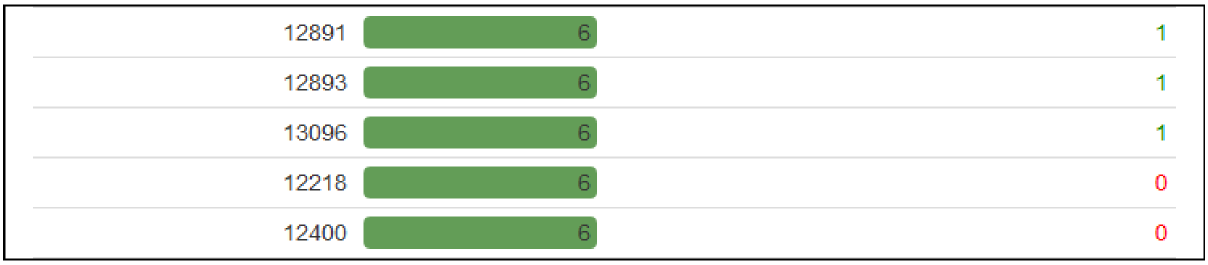
FIGURE 6.20: *Tab seven: Prediction and ranking of test set data (bottom).*

Recall that it was mentioned above that for the majority of cases a partition of the 1's and 0's of the weighted predictions column will occur within a single ranking level. An example of a case where this does not occur is shown in Figure 6.24.



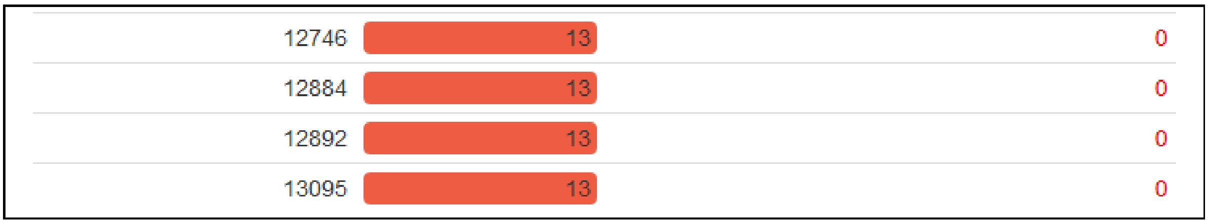
Alternative	Ranking	Weighted prediction
12005	1	1
12008	1	1
12020	1	1
12021	1	1

FIGURE 6.21: *Tab eight: Prediction and ranking of new alternatives (top).*



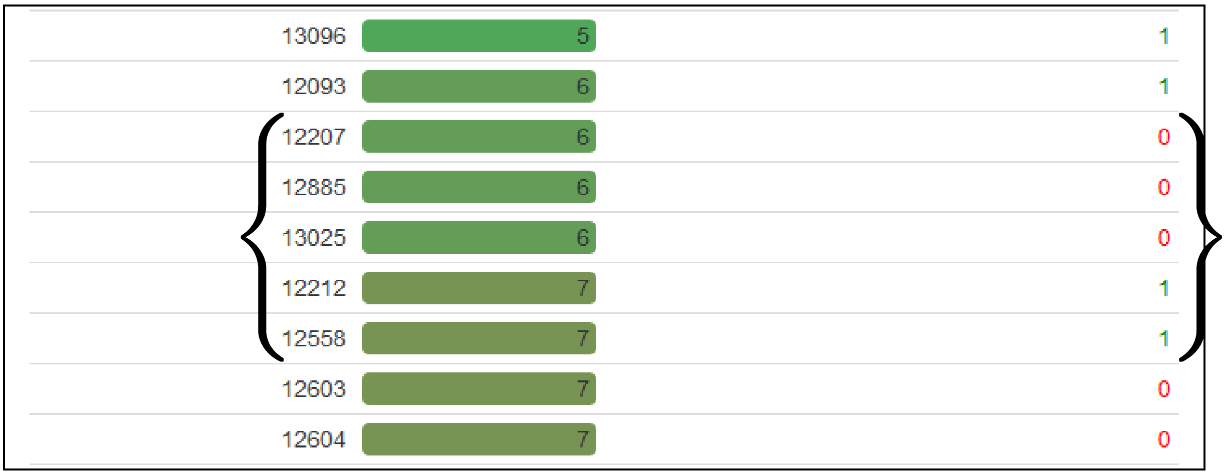
12891	6	1
12893	6	1
13096	6	1
12218	6	0
12400	6	0

FIGURE 6.22: *Tab eight: Prediction and ranking of new alternatives (middle).*



12746	13	0
12884	13	0
12892	13	0
13095	13	0

FIGURE 6.23: *Tab eight: Prediction and ranking of new alternatives (bottom).*



13096	5	1
12093	6	1
12207	6	0
12885	6	0
13025	6	0
12212	7	1
12558	7	1
12603	7	0
12604	7	0

FIGURE 6.24: *Discrepancy between the weighted prediction and ranking of new alternatives.*

The reason for the discrepancy is simply that the two steps of producing a weighted prediction and ranking of alternatives is determined from the same source information, that of the five prediction columns of the base models and their weightings. If the ranking followed the weighted prediction, or *visa versa*, more consensus could have been expected. Although the weighted prediction and the ranking source the same information, but occasionally produce conflicting information, the user should not see this as an error, but rather as a valid second opinion.

It is advocated that in such a scenario the observations for which the prediction and ranking levels do not agree should be grouped together and considered as borderline cases. For example, the five indicated observations (MyID 12207 to 12558) in Figure 6.24 should become the group of borderline alternatives, *i.e.* given a prediction of say 0.5 to indicate the fact that they are borderline, and a rank level 6.5. The remainder of the alternatives may remain as is. In the case that the weighted prediction and rank levels completely agree, it may be seen as two opinions in agreement, and thus a higher level of confidence in the results.

In addition to the working of the DSS showcased above, the DSS has various additional visual aids built into it to assist the user in successfully navigating the eight tabs. First, effective workflow is enforced in the DSS by partitioning the process into eight logical steps. The user can at any time see where he or she is in the overall process. An obvious concern is that the user might jump ahead or not complete the steps in the correct order. If a user attempts to skip any step by jumping ahead, he or she will be shown an error message indicating the need to first complete the previous steps. For example, the error message shown in Figure 6.25 is displayed to the user if he or she attempts to proceed to tab seven before completing all of the previous tabs. Similar messages will appear on any tab if the entries of the tabs prior to it have not yet been completed.

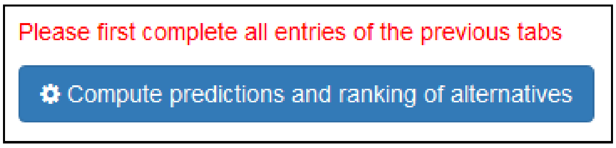
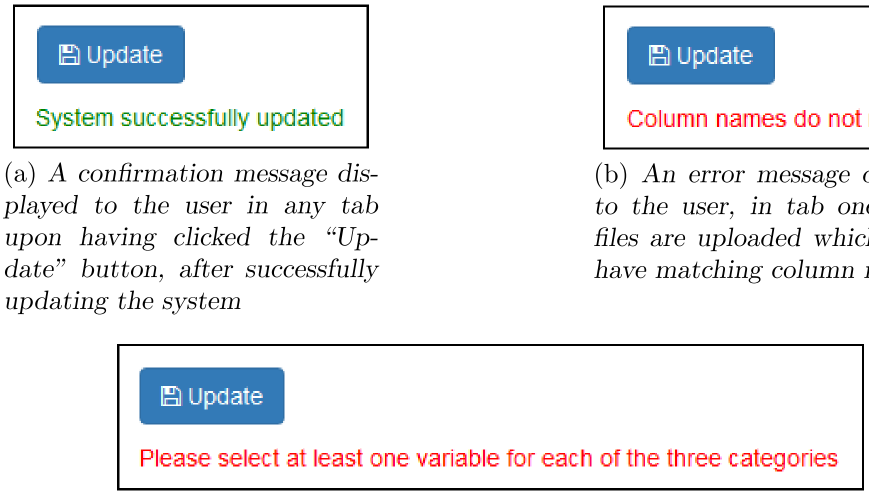

Please first complete all entries of the previous tabs

Figure 6.25: *Example of an error message presented to the user if he or she proceeds to a tab before completing all the entries in the previous tabs.*

The user is also provided with visual confirmation of valid entries and visual error messages, with reasons, for invalid entries. Three examples of such messages are shown in Figure 6.26.

Invalid entries into the system might also not be directly attributed to the result, but the user will nevertheless be notified and asked to correct such instances. An example of such an error occurs if the user were to select two independent variables that are perfectly correlated. The reason why this is considered an error is that the logistic regression model will not be able to fit such data. This error is only applicable to logistic regression, as the other four base models are able to accommodate perfectly correlated data. An example of the error message displayed to the user in this case is shown in Figure 6.27. In this example the user selected *Study region* and *Institution* as the two independent variables, which are obviously perfectly correlated, since each institution is located only in one study region (province).

Users of the DSS not only enjoy confirmation of correct entries, as explained above, they are also notified of their progress when using the system. For this purpose, a thin blue progress bar appears at the top of the DSS for any activity which takes longer than one second. An example of the progress bar displayed in tab seven is shown in Figure 6.28. The screen shot only captures one of the possible message outputs displayed below the progress bar as the messaging changes while the bar progresses from left to right. The other messages displayed to the user in tab seven are shown in Table 6.2.



System successfully updated

(a) *A confirmation message displayed to the user in any tab upon having clicked the "Update" button, after successfully updating the system*

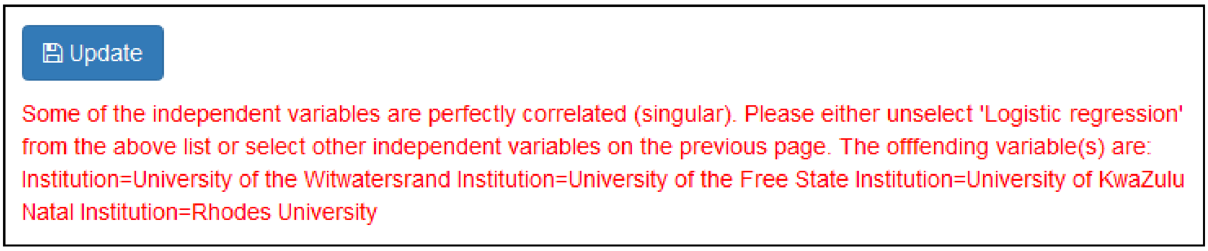
Column names do not match

(b) *An error message displayed to the user, in tab one, if two files are uploaded which do not have matching column names*

Please select at least one variable for each of the three categories

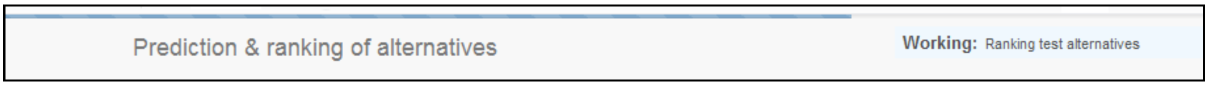
(c) *An error message shown to the user, in tab two, if he or she attempts to click the "Update" button before having selected at least one option from each of the three catogories*

FIGURE 6.26: *Three examples of confirmation and error messages displayed to the user while navigating the DSS.*



Some of the independent variables are perfectly correlated (singular). Please either unselect 'Logistic regression' from the above list or select other independent variables on the previous page. The offfending variable(s) are: Institution=University of the Witwatersrand Institution=University of the Free State Institution=University of KwaZulu Natal Institution=Rhodes University

FIGURE 6.27: *An example of the error message shown to the user in the case that singularity is found.*



Working: Ranking test alternatives

FIGURE 6.28: *An example of the progress bar in tab seven.*

TABLE 6.2: *Possible progress bar messages displayed to the user in tab seven.*

Percentage complete	Message
25%	Working: Constructing base models
45%	Working: Computing base model predictions
55%	Working: Combining base model predictions
70%	Working: Ranking test alternatives
85%	Working: Ranking new alternatives
95%	Working: Creating table

Not all the possible validation or error messages are shown in this discussion, neither are all of the other small functionalities which are intended to assist the user in navigating the system. It should, however, be clear from the above description that considerable effort was made to make the system as easy as possible to navigate by the user.

In addition, the system was designed keeping in mind the three golden rules mentioned in §6.5.4: place the user in control, make the user interface consistent and reduce the user's memory load. These guidelines were adhered to as follows.

First, the user was placed in control by giving him or her the ability to experiment with different alternatives. Within the bounds of what are considered valid entries, the user is at any time able to backtrack to the various analysis steps and start over by selecting a different combination of variables or new data set. At times the user is provided with information, such as indicators of model assumption validity, but he or she is still in control in the sense of being able to choose whether or not to continue with the analysis, by taking certain calculated risks when including specific base models in the analysis.

Secondly, it was a challenge to keep the GUI consistent in terms of visuals and the various steps of the process to be followed by the user as there are many different activities that the user must be able to perform. The author nevertheless kept the GUI as consistent as was deemed reasonable by placing sufficient emphasis on informing the user exactly where he or she is in the process and whether he or she is following the analysis steps correctly. This was achieved by making the top grey tab menu visible to the user at all times so as to provide a clear idea of where in the analysis the user finds him/herself. In the tabs that allowed it, the user was asked to "Update" the system and was given a confirmation of valid entries. These "Update" buttons and any other buttons which have to be clicked by the user were made the same colour blue. It should be clear to the user in each step of the analysis when he or she has done what is expected before continuing.

The third challenge was to design the system in such a manner that the user's memory load is not unnecessarily increased, *i.e.* the system was designed so that a balance was maintained between the two extremes of either inducing information overload on the part of the user and presenting an oversimplified interface. The eight tab menus also assisted in this as the user does not have to be concerned about his or her inputs on the previous tabs or of those to come, but is at any point provided with only the required information to make informed decisions on the current tab. Thus each of the tabs is designed so as to provide what is believed to be adequate information to make managerial decisions at any point during the analysis, without overwhelming the user with too much information, options or settings. An example of this is the following: Tab five was purposefully partitioned into six sub tabs, so that the user may quickly and easily investigate the summary table in the first tab, but the user is only shown all the exact results if he or she chooses to investigate further by making using of the sub-tabs. Another example is how the system in tab two determines which of the columns are suitable to be used as unique identifiers and as dependent variables. Although minor, this allows the user not to have to recall or figure out which fields are applicable to which of the three categories.

6.6 Quality assurance, verification and validation

Phases six (testing and maintaining the system) and seven (implementing and evaluating the system) of the SDLC are considered together in this section. It is highlighted how quality assurance methodologies were implemented throughout the entire process, and this is followed by an explanation of how testing and validation were carried out. A brief section on the system implementation is finally presented.

6.6.1 Quality assurance

The notion of quality assurance was discussed in §2.4.1. It was stated that there are three approaches that must be followed during any software development process in order to maintain quality assurance. They are: design the system using a top-down and modular approach, document the software using appropriate tools, and test and validate the system. During this system development in this study, the above-mentioned three approaches were considered in the following way.

The top-down and modular approaches that were used to analyse and design the system were discussed in §6.3 and §6.4. These sections highlighted how the author consistently considered how the interrelationships and interdependencies between the subsystems would fit into the larger system, in order to ensure quality and high-level integration. The top-down design approach enables the author to refrain from becoming caught up in minor details, but rather keep the bigger picture of what the system should achieve in mind.

The notion of quality assurance (appropriate software tools) was also indirectly covered in §6.3 and §6.4. As mentioned in §6.1, the first few phases of the SDLC were strictly followed in a waterfall manner, and this emphasis on documentation overlaps with the UML approach of §6.3 and the agile development approach of §6.4. Even more software documentation would have taken place by creation of an ERD for the database design, but as previously explained no such component is utilised in this study.

The third point above, concerned with the testing and validation of the system, is discussed in the following subsections.

6.6.2 System testing and verification

Testing (henceforth used interchangeably with the term *verification*), although tedious, is far more effective if performed from early development until after implementation, as opposed to only after implementation.

Four steps of testing were discussed in §2.4.2 and illustrated in Figure 2.8. Each of those steps are briefly discussed below, highlighting how each step was followed during this study.

The step of program testing by employing test data refers to the continual testing that occurs throughout the development of the system. The author tested new segments of code created by applying it to a small samples of test data. Any corrections required at this stage were then carried out.

While the step of testing with test data refers to isolated testing of small portions of code, the step of link testing with test data refers to testing all of those code portions together. Again using small samples of test data, the entire concept demonstrator was tested to make sure that all the segments work together as planned. Any corrections required at this stage were again affected.

Next, the entire system was tested by applying it to a large set of test data and documenting the entire process from the point of view of the GUI. During this step, the final user became involved with the system's building process and the documented GUI steps were presented to the user, as noted in Appendix A, so as to obtain feedback. Two important factors assessed were whether the output produced was understandable and sensible to the user, and whether the output produced was relevant and useful to the user. The feedback obtained from the final user, based on the above two factors, was vital and taken very seriously. Using this feedback, the author again made the necessary alterations and improvements to the system.

The step of full system testing with live data normally requires that the system be tested again in the same manner as described above, but with live data instead of test data. As live data were already used in the previous step, this step became superfluous and was not performed.

6.6.3 System validation

The difference between verification and validation was described in §2.4.3. Verification refers to the testing described in §2.4.2, which only verifies that what the systems analyst perceived to be the user requirements, acquired in the current context in §6.2, were satisfied. Thus, the testing performed in the previous section (§6.6.2) verified that the system performs as it was designed to do.

Validation, however, is the process of ensuring that the system that has been built, and successfully verified through testing, will actually meet the client's real needs, as opposed to only the needs that the systems analyst thought the client had when he was collecting the user requirements.

The author sat down with the client after the step of full system testing with test data was completed in order to discuss whether the system did, in fact, meet the client's real needs. As will be discussed in more detail later, the client was very satisfied with the system and the results it produced. The client stated that the system did meet their actual needs by providing, from input given by the client, relevant output that the client was able to interpret easily and use to make intelligent decisions.

At this point it may thus be concluded that all three approaches mentioned in §6.6.1, with regard to quality assurance, were indeed followed, and that as a result adequate steps were taken in this study to ensure quality throughout the development process of the DSS concept demonstrator.

6.6.4 Implementation and maintenance

The process of implementation (described in §2.4.4), is closely linked to the process of maintenance (discussed in §2.4.5). As previously mentioned, the DSS designed in this study is a concept demonstrator, and not a final product to be integrated into the main industry partner's computerised system. The steps of implementation and maintenance are thus not directly applicable to this study. As stated in §6.1, however, the DSS was designed in such a way as to allow for a fruitful discussion to take place with an external programming company who should be able to implement the DSS and integrate it with the systems of the main industry partner, should the main industry partner desire it. This discussion took place on the 4th of August 2016, as indicated on the communication registry of Appendix A. The client's opinion as to the feasibility of the proposed DSS was positive as they believe that such a system will add value to a bursary provider, but unfortunately it falls outside the scope of their current focus and hence they do not currently have the capacity to take the project on. Their exact feedback was as follows [245]:

> *"The student demonstrated an application to predict the outcome of whether a bursary applicant will succeed at their studies or not, based on known data from historical beneficiaries. Although this is not within our development scope for DevMan at this point, it is an interesting tool that our bursary clients will likely find useful in the recruitment and selection phase of bursary applications."*

Despite this the main industry partner stated that they are satisfied with the current documentation and video demonstrations that they received and that they will in the future be able to approach an alternative programming company, more specialised in such systems, to create a similar system for them.

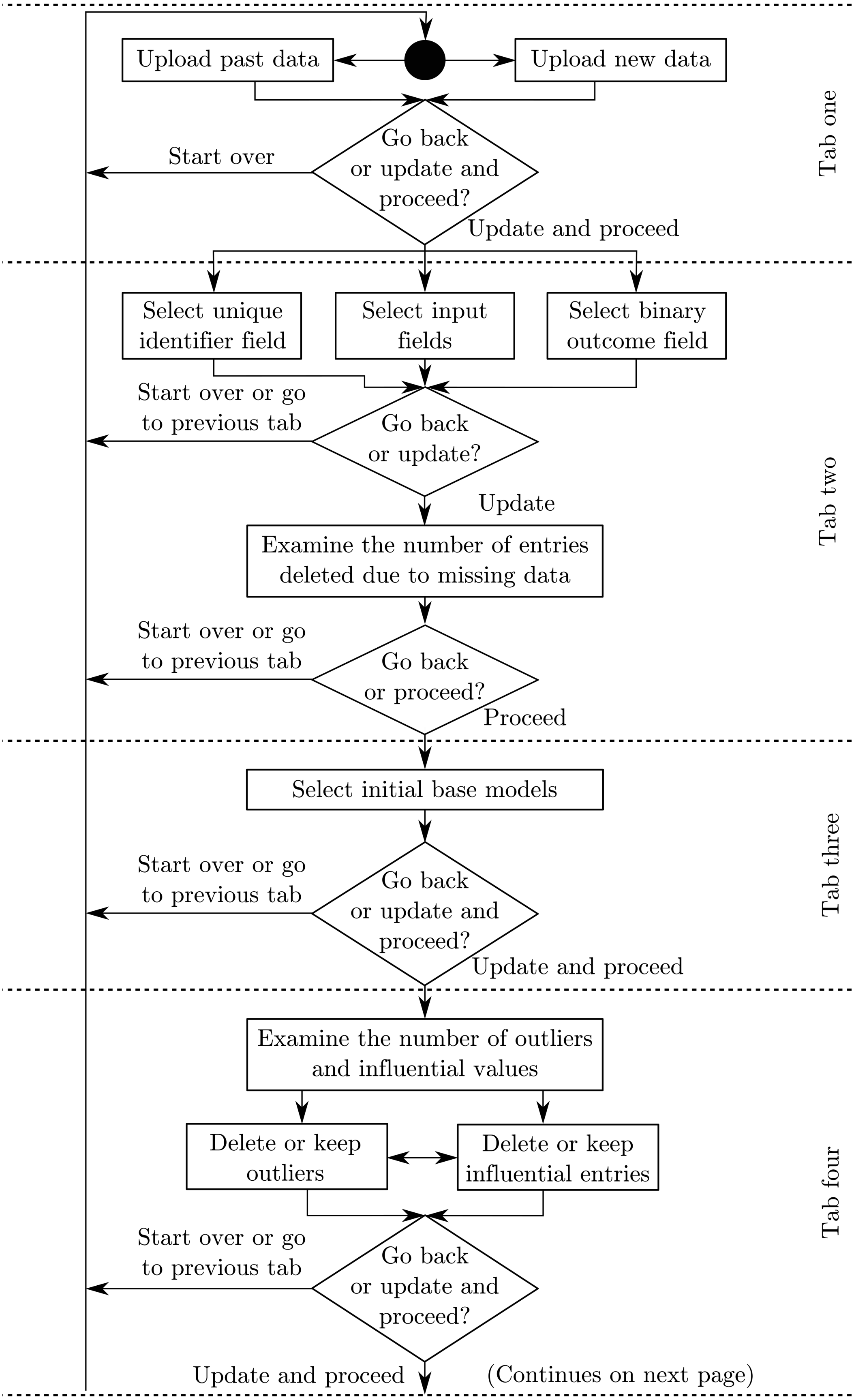

Upload past data
Upload new data
Go back or update and proceed?
Start over
Update and proceed
Tab one
Select unique identifier field
Select input fields
Select binary outcome field
Start over or go to previous tab
Go back or update?
Update
Examine the number of entries deleted due to missing data
Start over or go to previous tab
Go back or proceed?
Proceed
Tab two
Select initial base models
Start over or go to previous tab
Go back or update and proceed?
Update and proceed
Tab three
Examine the number of outliers and influential values
Delete or keep outliers
Delete or keep influential entries
Start over or go to previous tab
Go back or update and proceed?
Update and proceed
(Continues on next page)
Tab four

FIGURE 6.29: *First GUI flow from the perspective of the user.*

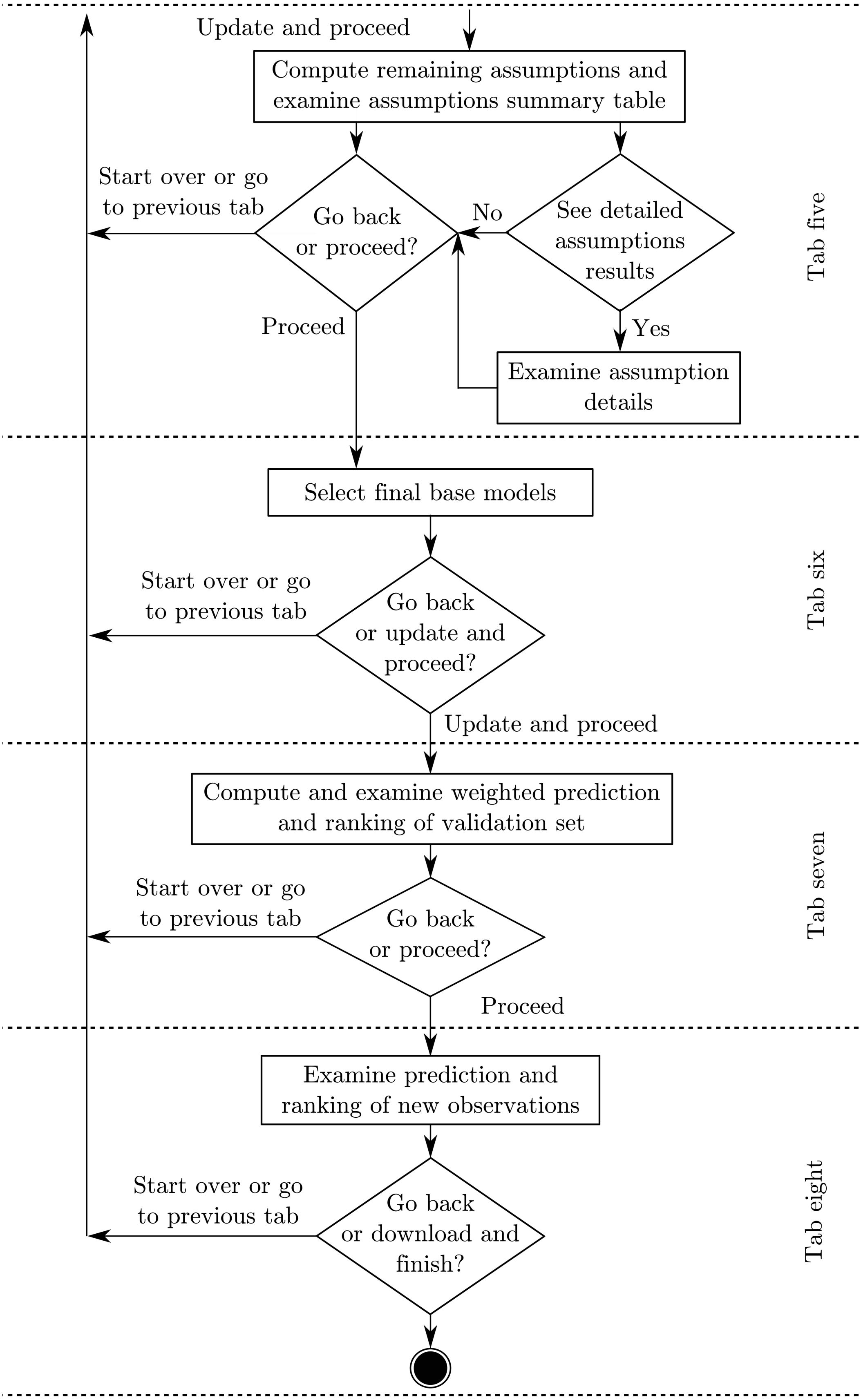

Update and proceed
Compute remaining assumptions and examine assumptions summary table
Start over or go to previous tab
Go back or proceed?
No
See detailed assumptions results
Tab five
Proceed
Yes
Examine assumption details
Select final base models
Start over or go to previous tab
Go back or update and proceed?
Tab six
Update and proceed
Compute and examine weighted prediction and ranking of validation set
Start over or go to previous tab
Go back or proceed?
Tab seven
Proceed
Examine prediction and ranking of new observations
Start over or go to previous tab
Go back or download and finish?
Tab eight

FIGURE 6.30: *Second GUI flow from the perspective of the user.*

6.7 Chapter summary

The main purpose of this chapter was to present and discuss the design of a concept demonstrator of the DSS of the previous chapter. The SDLC was used as the foundation of the development process, and each of its phases was performed according to a specific SDM style. It was first described how each SDM style was incorporated into the SDLC. Thereafter, the specifics of each of the seven SDLC phases were presented.

First, the objectives of the study were referred to and the user requirements obtained from discussions with the main industry partner were presented. Thereafter, an analysis of the system needs took place, graphically illustrating the current state of the main industry partner's system by means of use case and activity diagrams. Next, a discussion took place on the design of the system. Thereafter, topics applicable to the building of the concept demonstrator were considered, including discussions on data preparation, data assumptions, software considerations, framework element selection, and GUI considerations. A GUI system walk-through of the DSS concept demonstrator was finally provided, using real data. The process of using the GUI from the perspective of the user is summarised in the flow diagram of Figures 6.29–6.30. Finally, the measures taken to achieve quality assurance (in the context of system verification, system validation, and system implementation) of the concept demonstrator during its design and development was described.

Part III

Case studies

CHAPTER 7

Case study data presentation and preparation

Contents

7.1 Target population 146
7.2 Discussion of the main industry partner sample data 146
7.3 Case Study A aim and variable selection 152
7.4 Case Study B aim and variable selection 154
7.5 Discussion of the secondary NGO sample data 158
7.6 Case Study C aim and variable selection 160
7.7 Case Study D aim 162
7.8 Chapter summary 162

The purpose of this chapter is to present the data obtained from two industry partners and to discuss the aims of four case studies to be performed in respect of these data. First, the target population of tertiary students relevant to these case studies is elucidated and briefly discussed. The remainder of the sections are applicable to the four case studies for which brief descriptions are as follows:

Case Study A This case study utilises sample data provided by the main industry partner. The aim of the case study is to investigate whether the three score variables, described in §7.2.6, collected by the main industry partner during the application process, may be used to predict academic success of applicants reliably.

Case Study B This second case study is concerned with an overlapping sample to the one used in the Case Study A. For this case study, rather than having to investigate specific predefined variables, the aim is to assess all variables in the sample for variable importance and other interesting trends.

Case Study C For this case study, data obtained from the secondary industry partner were utilised. In a similar fashion to Case Study B, the aim of this case study is to investigate all variables in the sample for variable importance and other interesting trends.

Case Study D The fourth and final case study is concerned with assessing the performance of the DSS concept demonstrator of Chapter 6 in the context of the data of Case Studies B and C.

These case studies are incorporated in the remainder of this chapter in the following manner. In the second section of this chapter, the data to be used in Case Studies A and B, as provided by the main industry partner, are presented, and this is followed by two sections in which the specific aims, variable selections and data cleaning processes of Case Studies A and B are discussed, respectively. In a similar fashion, the data to be used in Case Study C, as provided by the secondary industry partner, are presented in the next section and this is followed by a section elaborating on the aim, variable selection and data cleaning process of Case Study C. Next, a section is dedicated to discussing the aim of Case Study D. Finally, the chapter closes with a brief summary of its contents.

7.1 Target population

As was mentioned in Chapter 1, data relating to underprivileged tertiary students are analysed in this thesis. Figure 7.1 contains a graphical representation of all underprivileged tertiary students from around the world. The target population for the study is underprivileged South African tertiary students. These students may be seen as all those who fall within the "Lower Class" tab of Figure 7.1. To facilitate the study, data samples of students who fall within this target population were provided by the main industry partner and by a secondary industry partner. In theory, these samples are a representation of the target population.

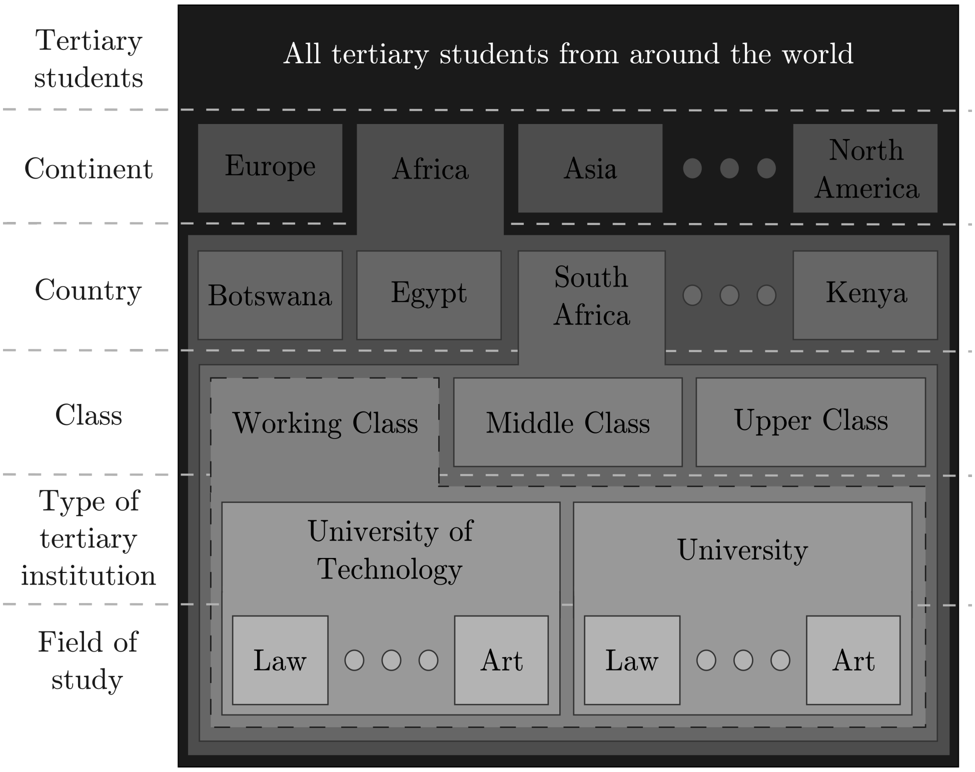


FIGURE 7.1: *Target population applicable to this study.*

7.2 Discussion of the main industry partner sample data

The main industry partner, founded in 2001, facilitates the allocation of bursaries to individuals originating from poor, rural communities within South Africa. Their sponsors, who provide financial donations, include a wide variety of both private and governmental bodies.

A description of the sample data provided by the main industry partner is provided in this section. The variable names of the data categories provided by the industry partner are shown in Table 7.1. Each of the fields in this table are expanded on in the subsections that follow. Two overlapping sets of this sample data will be used for Case Studies A and B in the following chapter.

Table 7.1: *Fields in the sample data used in Case Studies A and B.*

	Number	Data field	Data type
Personal information	1	Gender	Dichotomous
	2	Race	Multichotomous
	3	Family income	Ratio
	4	Number of members in family	Ratio
	5	Family income per member	Ratio
Geographical	6	Home/Source region	Multichotomous
	7	Tertiary study region	Multichotomous
Academic administration	8	High school institution	Multichotomous
	9	Tertiary institution	Multichotomous
	10	Start year	Discrete
	11	Expected end year	Discrete
	12	Actual end year	Discrete
	13	Throughput	Discrete
	14	Variance	Discrete
	15	Academic status	Ordinal
	16	Study field	Multichotomous
	17	Qualification type	Multichotomous
	18	Qualification	Multichotomous
High school academics	19	Grade 12 Nov average	Discrete
	20	Grade 12 Nov subject names	Nominal
	21	Grade 12 Nov subject marks	Discrete
	22	Grade 12 June average	Discrete
	23	Grade 12 June subject names	Nominal
	24	Grade 12 June subject marks	Discrete
	25	Grade 11 Nov average	Discrete
	26	Grade 11 Nov subject names	Nominal
	27	Grade 11 Nov subject marks	Discrete
Tertiary academics	28	Tertiary semester average	Discrete
	29	Tertiary subject names	Nominal
	30	Tertiary subject marks	Discrete
Scores	31	Form score	Discrete
	32	Interview score	Discrete
	33	Recommendation	Discrete

7.2.1 Personal information

Five of the data fields relate to personal information of the students, namely *Gender*, *Race*, *Family income*, *Number of members in family* and *Family income per member*.

The fields of *Gender* and *Race* indicate respectively the gender and race of the students. The available options for these fields may be seen in Tables 7.2–7.3.

Table 7.2: *Gender options applicable to Case Studies A and B.*

Gender options	
1	Male
2	Female

Table 7.3: *Race options applicable to Case Studies A and B.*

Race options	
1	Black
2	Coloured

Family income represents the total income of the family of which the student is a member, while *Number of members in family* indicates the size of that family or the number of people supported by the family income. The field *Family income per member* refers to the average income per family member of the household from which the student originates, and may be calculated by dividing the family income by the number of members in the family.

7.2.2 Geographical information

The two fields related to the geographical information of the students are the *Home/Source region* and *Tertiary study region* fields.

The *Home/Source region* field indicates the province and region, or city, within that province, from which the students originate. The *Tertiary study region* is only the province, but not the region or city within that province, in which the students' tertiary institutions are situated. For both these geographical fields the possible entries are one of the nine provinces of South Africa. These are listed in Table 7.4.

Table 7.4: *South African provinces applicable to Case Studies A and B.*

South African provinces					
1	Eastern Cape	4	KwaZulu-Natal	7	Northern Cape
2	Free State	5	Limpopo	8	North West
3	Gauteng	6	Mpumalanga	9	Western Cape

7.2.3 Academic administration information

The field *High school institution* refers to the high school at which the student completed Grade 12. Within the sample, 492 different high schools were identified.

The field *tertiary institution* represents the name of the Tertiary institutions the students plan to attend. The sixteen tertiary institutions currently supported by the main industry partner are listed in Table 7.5.

Table 7.5: *Tertiary institutions supported by the main industry partner and applicable to Case Studies A and B.*

Tertiary institutions supported by the main industry partner			
1	Cape Peninsula University of Technology	9	Tshwane University of Technology
2	Central University of Technology	10	University of Cape Town
3	Durban University of Technology	11	University of Johannesburg
4	Nelson Mandela Metropolitan University	12	University of KwaZulu-Natal
5	North West University	13	University of Pretoria
6	Rhodes University	14	University of the Free State
7	Sefako Makgatho Health Sciences University	15	University of the Western Cape
8	Stellenbosch University	16	University of the Witwatersrand

The *Start year* field indicates the year in which a student started his or her tertiary studies of the last *attempted* course (it may not have been their first course). *Expected end year* represents the year in which a student is expected to graduate, and *Actual end year* the year in which he or she did, in fact, graduate.

From these three dates is it possible to calculate a student's *Throughput*, which represents the number of years they studied. It is also possible to obtain the *Variance* for each student, which indicates by how many years a student studies more or fewer than was expected of them. A variance of zero indicates that a student studied for the originally expected number of years, while a positive or negative variance indicates that he or she studied longer or shorter, respectively, than originally expected. A student is typically only assigned a negative variance in the case that they withdrew before obtaining the intended qualification.

The *Academic status* field indicates the students' current academic status. A student is assigned one of two main academic statuses: either *Graduated* or *Withdrawn*. *Graduated* indicates that the student has successfully completed his or her tertiary degree. *Withdrawn* indicates that a student has withdrawn from their tertiary studies. The two main statuses of *Graduated* and *Withdrawn*, along with the possible reasons for withdrawal, are shown in Table 7.6.

TABLE 7.6: *Academic statuses applicable to Case Studies A and B.*

Academic status			
1	Graduated		
2	Withdrawn	2.6	Other
2.1	Academic exclusion	2.7	Other funding
2.2	Cancelled studies	2.8	Poor academics
2.3	Deceased	2.9	Studies postponed
2.4	Medical	2.10	Support declined
2.5	Non-compliance	2.11	Support suspended

Elucidations of each of the possible reasons in Table 7.6 for withdrawal are as follows:

Academic exclusion: The student was prohibited by the tertiary institution from continuing with their studies due to their poor academic record.

Cancelled studies: The student cancelled his or her studies by their own choice.

Deceased: The student passed away while in the process of pursuing their qualification.

Medical: The student had to withdraw due to medical reasons.

Non-compliance: The student did not comply with the industry partner's non-academic bursary requirement, such as attending monthly meetings with his or her advisor.

Other: The student withdrew for an unknown reason.

Other funding: As the industry partner only provides funding to students who have no other bursaries, a student has to withdraw if he or she receives other funding.

Poor academics: Funding was withdrawn by the industry partner due to the poor tertiary academic record of the student.

Studies postponed: The student has suspended their studies for whatever reason.

Support declined: Similar to poor academics, the industry partner chose to withdraw funding from the student for academic reasons.

Support suspended: Also similar to poor academics, the industry partner chose to withdraw funding from the student for academic reasons.

Study field refers to the academic field in which the student engages in tertiary studies. The twelve study fields applicable to this data field are shown in Table 7.7.

The *Qualification type* field indicates the type of qualification the student either unsuccessfully attempted to or successfully obtained. The three possible types of qualifications are listed in Table 7.8.

Table 7.7: *Study fields applicable to Case Studies A and B.*

Study field			
1	Arts	7	Humanities
2	Built Environment	8	Law
3	Business Management	9	Management
4	Commerce	10	Medical
5	Education	11	Science
6	Engineering	12	Technology

Table 7.8: *Qualification types applicable to Case Studies A and B.*

Qualification type	
1	Degree
2	Extended degree
3	National diploma

The qualification of a student refers to the exact qualification he or she attempted to obtain or did obtain (*e.g.* BSc Chemical Science or Diploma of Graphic Design). There are a total of 194 different qualifications listed in the sample.

7.2.4 High school academic information

For each of the three high school examination periods of Grade 11 November, Grade 12 June, and Grade 12 November, the sample contains three fields, namely *Averages*, *Subject names*, and *Subject marks*, each of which are discussed below.

The fields of *Grade 12 Nov average*, *Grade 12 June average*, and *Grade 11 Nov average* contain the examination averages obtained by the student for those respective examinations. These are recorded in percentile form and thus range from 0% to 100%. The high school subject *Life orientation* is not included in the calculation of these averages, since it is commonly not taken into consideration when determining the average marks of students by prospective employers or bursary providers.

The data also contain fields indicating the names of the subjects taken by students during their high school careers. Sixty one different high school subjects were identified within the sample. Only sixteen of these were taken by at least a hundred of the students (*i.e.* by at least 9.1% of the students). The names of these sixteen subjects, along with counts of the number of students who took the subjects and the percentages of the total number of students who took the subjects are shown in Table 7.9. As some students take more than one of these subjects, the percentages do not add up to 100%. The marks achieved by students for their various subjects are recorded in percentile form.

7.2.5 Tertiary academic information

For each of the bi-annual semesters during which a student attended a tertiary institution, the data contains three fields, namely the *Semester average*, *Subject names*, and *Subject marks.*

The *Semester average* is the average that a student obtained for all his or her subjects during that semester. The *Subject names* field contains the names of the student's subjects of that semester and the *Subject mark* field contains the marks obtained by the student for each of their subjects. Both the *Semester averages* and *Subject marks* are recorded in percentile format.

Table 7.9: *High school subjects applicable to Case Studies A and B.*

Subject name	Count	Percentage of total
Accounting	276	25.0
Afrikaans (1st)	106	9.6
Afrikaans (2nd)	188	17.1
Agricultural science	121	11.0
Business economics	279	25.3
Economics	210	19.1
English (2nd)	903	81.9
Geography	376	34.1
History	111	10.1
IsiXhosa (1st)	153	13.9
IsiZulu (1st)	275	25.0
Life sciences	623	56.5
Mathematics	730	66.2
Maths Literacy	133	12.1
Physical science	544	49.4
Setswana (1st)	124	11.3

7.2.6 Score information

The bursary application process of the main industry partner was described in §6.3, including how the three scores *Form score*, *Interview score*, and *Recommendation* are collected. More details on these scores are provided in this section.

The *Form score* comprises four components, each of which is rated on a scale of 1 to 3. The total form score is calculated by combining the scores of the four components. A student may thus receive a minimum of 4 and maximum of 12 as a form score.

The first component, *Form completion*, evaluates the completeness of the bursary application form as filled in by the student. The student's ability to correctly understand and complete all the sections in the form in an accurate manner is seen as a potential positive indicator. It thus counts against any student who does not complete a section of the form. The second, *Course motivation*, refers to how well a student understands the nature and subjects of the course for which he or she is applying. One of the questions on the form specifically asks the student for an outline of his or her understanding of the course and how he or she expects to apply the knowledge gained from the degree in the workplace. The next component, *English score*, is not related to a student's high school English mark, but rather the quality of English used in completing the form. The *Extra-mural score* is the final component and is assigned based on the level of extra-curricular activities in which the student participates as well as their general involvement in the home community.

The *Interview score* is assigned to a student based on how well he or she does during the application interview. During the interview the student is evaluated on a scale of 1 to 5, with 1 indicating a weak response and 5 an excellent one, within five different sections. Each student will thus receive an interview score of between 5 and 25 (inclusive). Different types of questions are asked in each of the five sections. The five sections are: knowledge of field of study; preparation for study; independence and coping skills; self-motivation and determination; and self-confidence and interest in the bursary.

Each student is also assigned a recommendation score of 1 to 5 by the recruiter based on his or her personal impression of the student. The criteria for assigning the scores are listed in Table 7.10.

Table 7.10: *Recommendation score criteria applicable to Case Study A.*

Score	Criteria
1	Not recommended
2 & 3	Recommended with reservation
4	Recommended
5	Highly recommended

7.3 Case Study A aim and variable selection

Two case studies are carried out in the next chapter of this thesis, based on different, but overlapping, samples of the main industry partner's data, as described in §7.2. This section relates to the first of these two case studies, Case Study A, and the next section to the second case study, Case Study B.

The aim of Case Study A is to answer a very specific question posed by the main industry partner, namely whether the scores described in §7.2.6 and collected by them during the application process may be used to predict academic success of applicants reliably. The three variables related to these scores are thus the primary focus of this case study.

In order to achieve the objective of the case study, specific dependent and independent variables have to be selected, and the sample data have to be adequately prepared. A discussion on these activities now follow.

7.3.1 Dependent variable selection for Case Study A

For the purpose of this case study, the dependent variable was chosen such that those students who graduated successfully from their respective tertiary institutions may be identified. The dependent variable was therefore chosen as the *Academic status* of a student. The possible academic status that a student may have was presented in Table 7.6. Those students in the sample data with an *Academic status* of *Graduated* were considered academically successful and were therefore assigned a value 1 for the dependent variable.

Students with an *Academic status* of *Cancelled studies*, *Deceased*, *Medical*, *Other*, or *Studies postponed* were not considered, because the reason for withdrawal is either unknown or was due to circumstances outside the students' control. Although *Non-compliance* by the student is not a positive phenomenon, it is a non-academic issue and students who did not comply were therefore also not considered during the analysis. The situation where a student received other funding is considered by the industry partner as a success, but since the final academic outcome of those students (graduated or not) is unknown, they were also excluded from consideration. The only remaining academic statuses due to withdrawal are *Academic exclusion*, *Poor academics*, *Support declined*, and *Support suspended* which were considered indications of academic failure and were thus assigned a value of 0 for the dependent variable.

Due to the nature of the study, only students who had either completed or withdrawn from their tertiary studies were considered and no student currently still pursuing their tertiary studies was considered.

7.3.2 Independent variables for Case Study A

Although the focus in this case study was on the three score variables, it was required to test these variables against others for comparison purposes. The most logical choice was the three other variables currently considered by the main industry partner during the academic application process, namely the variables of *Grade 11 Nov average*, *Grade 12 June average*, and *Grade 12 Nov average*.

In addition, the variable of *Qualification type* was also selected as an independent variable so as to have even more variables to compare the score variables against. Since data are available for the vast majority of observations in the sample for this variable, its inclusion barely reduces the sample size. No other independent variables were considered as doing so would significantly reduce the sample size due to the large number of missing values in the fields of other possible independent variables.

7.3.3 Data preparation for Case Study A

The relevant data set was obtained from the industry partner who provided the data in a tabular format within `Excel.xlsx` formatted files. The data fields contained within the data set were discussed in §7.2.

The various sheets of the file needed to be integrated into a single sheet/table and reworked in such a manner so as to allow for only one student's information to be present in a single row. Finally, the file was saved in an `Excel.csv` format as to be used for input in the relevant data analysis software.

The data set provided by the main industry partner for the purpose of this case study contained all the variables listed in §7.2 for 244 observations. Of these 244 students, 55 had an *Academic status* which was not one of the desired types required in order to be able to be labelled either successful or unsuccessful. After the removal of these 55 students, only 189 remained in the sample.

Thereafter, erroneous data records were identified and corrected within the data set. Data entries were found to be erroneous either because they contained missing values, valid errors, and/or invalid errors. These types of erroneous data were discussed in §2.2.4. Each of these erroneous entries may be corrected by having the specific entry removed, replaced, corrected, or left as is. The number of cases of each type of erroneous data entry was identified within the sample and the corrective strategy chosen for each type is shown in Table 7.11. Types of corrective strategy not applicable to specific types of erroneous data are marked as "N.A." in the table.

TABLE 7.11: *Possible types of erroneous data and applicable corrective strategies for the sample data of Case Study A.*

	Number	Chosen corrective strategy:			
	of cases	Removed	Replaced	Corrected	Left as is
Missing values	148	✓		N.A.	N.A.
Valid error values	3	✓			N.A.
Invalid error values	1	✓			N.A.

Specifically which independent variables could be used in the analysis was dictated by how many of each type of erroneous entries were observed, as shown in Table 7.12. As many of the data entries containing erroneous values overlapped, only 80 entries had to be removed from the sample, therefore resulting in a final sample of 109 data records.

The data were only assessed for outliers once all the missing, valid error, and invalid error values had been identified and corrected (*i.e.* only the 109 remaining students were thus assessed). The data set was searched for potential outliers by using the proximity matrix produced during the implementation of the statistical technique of *random forests*, discussed in §3.7. An outlier detection value, specific to random forests, was determined for each entry of the sample using the method outlined in §3.7.2. The results are shown in Figure 7.2. As previously stated, any entry with an outlier detection value exceeding 10 may be considered a possible outlier. In the case of this data set, the highest outlier detection value calculated was just above 6 and it was therefore deemed that no significant outliers are present in the data set.

Table 7.12: *Numbers of each independent variable corrected due to erroneous values present in the sample data of Case Study A.*

	Missing values	Valid error values	Invalid error values
Academic administration			
Qualification type	2		
High school academics			
Grade 11 Nov average	55		
Grade 12 June average	56		
Grade 12 Nov average	14		
Scores			
Form score	9		
Interview score	8	2	1
Recommendation	2	1	

At the same time the data were analysed for influential observations using Cook's Distance as outlined in §2.2.4. As none of the 109 observations were identified as being influential values, it was not necessary to consider their removal.

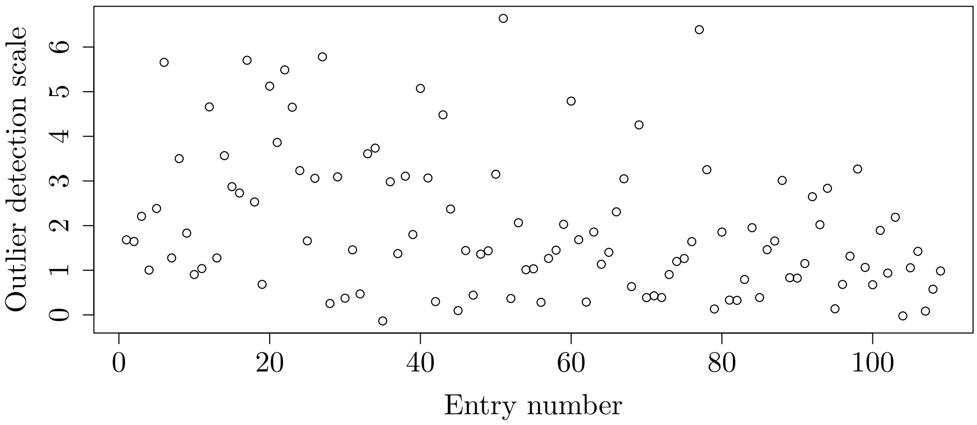


Figure 7.2: *Outlier detection performed on the sample data of Case Study A.*

7.4 Case Study B aim and variable selection

In the case study alluded to in the previous section, Case study A, the main industry partner posed a question related to the three specific score variables collected by them during the bursary application process. In a follow-up case study, all remaining variables will be assessed with the aim of identifying those individual variables, as well as the largest combination of variables, that produce the highest variable importance and significance scores in terms of the expected success of students during their tertiary studies. In addition, any other interesting trends within the data will also be investigated.

7.4.1 Dependent and independent variable selection for Case Study B

In order to achieve the objective of this case study, suitable dependent and independent variables have to be selected. The same dependent variable as that identified in §7.3.1, namely *Academic status* (variable 15 in Table 7.1), is again used in this case study.

All remaining variables listed in Table 7.1, except variables 10–14 and 28–33, are eligible as possible independent variables. All data that will only be available for an applicant once they have started their studies were removed from consideration as no data of this kind will be available to the industry partner at the time that students commence their tertiary studies. The eight variables excluded from consideration for this reason are variables 10, 11, 12, 13, 14, 28, 29, and 30 in Table 7.1. The three score variables numbered 31, 32, and 33 in Table 7.1 are also not considered since the second data sample provided by the main industry partner did not contain these fields due to a lack of score data.

The following twenty one data fields are therefore considered as possible independent variables in the follow-up case study: variables 1–9 and 16–27 in Table 7.1.

7.4.2 Data preparation for Case Study B

The process of data preparation within the context of the DDS framework was presented in Figure 5.2, and builds on the concepts of data preparation discussed in §2.2. For the sample data at hand, the steps of data gathering, data integration, and data extraction have all been described in §7.3.3.

Although only sixteen institutions are currently listed in Table 7.5, an additional twenty six tertiary institutions appear within the data sample. These are listed in Table 7.13.

Table 7.13: *Additionally listed tertiary institutions in the sample data of Case Studies A and B.*

Additionally listed tertiary institutions			
1	Border Technikon	14	University of Durban-Westville
2	Cape Town Technikon	15	University of Fort Hare
3	Durban Institute of Technology	16	University of Limpopo
4	Eastern Cape Technikon	17	University of Natal
5	Medical University of South Africa	18	University of Port Elizabeth
6	Peninsula Technikon	19	University of Potchefstroom
7	Port Elizabeth Technikon	20	University of Transkei
8	Randse Afrikaanse University	21	University of Venda
9	Technikon Mangosuthu	22	University of Zululand
10	Technikon North West	23	University of South Africa
11	Technikon Northern Gauteng	24	Vaal Triangle Technikon
12	Technikon of the Orange Free State	25	Vaal University of Technology
13	Technikon Pretoria	26	Wits Technikon

These additional institutions appear within the data due to tertiary institution merges, name changes, or due to subsequent decisions by the main industry partner to discontinue supporting certain institutions. The relevant changes were made in the data so as to reflect only the most updated names of the tertiary institutions currently supported by the main industry partner. Specific changes were made due to the following:

- In 2005, the *Randse Afrikaanse University* and *Wits Technikon* merged to form the *University of Johannesburg.*
- Both the *University of Durban-Westville* and the *University of Natal* were incorporated into the *University of KwaZulu-Natal.*
- The *Eastern Cape Technikon*, the *University of Transkei*, and the *Border Technikon* were combined to form *Walter Sisulu University* in 2005.
- *Nelson Mandela Metropolitan University* was formed by combining of the *University of Port Elizabeth* and *Port Elizabeth Technikon.*
- The merger of *Cape Town Technikon* and *Peninsula Technikon* in 2005 resulted in the creation of *Cape Peninsula University of Technology.*
- The *Durban Institute of Technology* had its name changed to the *Durban University of Technology.*
- The *Tshwane University of Technology* was formed in 2004 by the combination of the *Technikon North West*, the *Technikon Northern Gauteng*, and the *Technikon Pretoria.*
- In 2004, *North West University* came about as a merger of the former *Potchefstroom University for Christian Higher Education* and the former *University of North West* (formerly known as the *University of Bophuthatswana*).
- The *Vaal University of Technology* was previously known as the*Vaal Triangle Technikon.*

In all of the above cases, the tertiary institution names were changed in the data sample so as to appropriately reflect the new names of the institutions.

The most intricate change, however, was the following: *Sefako Makgatho Health Sciences University* is now a separate entity, which was the old *Medical University of South Africa.* The *Medical University of South Africa* was originally one of the campuses of the *University of Limpopo* (their medical campus). It was thus required to locate all the students in the data listed as having studied at the *Medical University of South Africa* or who studied a medicine-related field, such as Radiography or Optometry, at the *University of Limpopo* and change their tertiary institution field to *Sefako Makgatho Health Sciences University.* The remainder of the students with a tertiary institution label of *University of Limpopo* were left as is.

Support to nine of the institutions listed in Table 7.13 has been discontinued by the main industry partner. These nine institutions are *Technikon Mangosuthu*, *Technikon of the Orange Free State*, *University of Fort Hare*, *University of Limpopo*, *University of Venda*, *University of Zululand*, *University of South Africa*, *Vaal University of Technology*, and *Walter Sisulu University.* The information related to these institutions will not be useful again when predictions are carried out based on a new data set as any new data set will not contain any information related to these institutions. The sample provided by the main industry partner for Case Study B contains 1 445 students. For this sample the combined number of students listed as affiliated with these nine institutions are a mere 87 (that is, 6% of the sample). This is a minor adjustment in comparison to the sixteen institutions still supported, which represent the remaining 94% of the students. It is also possible that during the learning process of the base models, all the students listed at one of these institutions may well fall into the test set and none into the training set. This may result in undesirable computational results in the sense that the DSS concept demonstrator may not be able to predict the outcome of a student attending that specific tertiary institution.

For this reason it was decided to remove the tertiary institutions of these 87 students from the sample data. Note that their entire data records were not removed — only the field relating to their tertiary institutions.

Of the 1 445 students, 344 students had to be removed due to not having been associated with one of the desired dependent variable labels, as mentioned in §7.3.3. The numbers of students who were excluded as a result of incorrect or missing dependent variable labels are shown in Table 7.14. The dependent label "Missing entry" points to students who had a blank *Academic status* entry. This resulted in a sample of 1 101 students.

Table 7.14: *Numbers of students deleted (cumulative) due to incorrect or missing dependent variable labels in the sample data for Case Study B.*

Missing or incorrect dependent label	Number deleted	Missing or incorrect dependent label	Number deleted
Cancelled studies	92	Other	9
Deceased	1	Other funding	124
Medical	5	Studies postponed	5
Non-compliance	61	Missing entry	47

Erroneous data records were also sought and corrected within the data sample, as follows. Data entries were found to be erroneous either because they contained valid error values and/or invalid error values. As mentioned in the previous section, each of these types of erroneous entries may be corrected by having the specific entry removed, replaced, or corrected. The number of cases of each type of erroneous data entry was identified within the sample and the corrective strategy chosen for each type is shown in Table 7.15.

The number of each type of erroneous entry observed for each applicable independent variable is shown in Table 7.16. Unlike in the data for Case Study A, the "Remove" option does not relate to removing the entire observation, but rather only the datum in that specific field of the particular student. This resulted in none of the observations being removed completely and so the sample still contained 1 101 students after the cleaning process.

Table 7.15: *Identification of valid and invalid error values and applicable corrective strategies in the sample data for Case Study B.*

	Number	Chosen corrective strategy:		
	of cases	Removed	Replaced	Corrected
Valid error values	305	✓(297)		✓(8)
Invalid error values	1	✓(1)		

Note that the two types of erroneous entries known as outliers and influential observation values and missing values are not considered at this stage, for the following reasons. The lack of outliers is an assumption of some base models. Thus when the assumptions are validated, the user is given the power to assess the outlier score of each observation in the sample and choose a remedial action he or she sees fit. Missing values are also not considered yet since, unlike in the data for Case Study A (described in §7.3), the independent variables to be selected by the user are not yet known and it would be unwise to delete an entire observation due to a missing entry for one of the fields which might not be selected as an independent variable. As the user selects the independent variables in the DSS, they is notified of the number of observations removed due the choice. The complete list of independent variables available to the user during Case Study B, as well as the numbers of entries missing from each, may be seen in Table 7.17.

TABLE 7.16: *Numbers of each independent variable corrected due to erroneous values present in the sample data of Case Study B.*

	Valid error values	Invalid error values
Personal information		
Family income	152	1
Number of members in family	6	
Income per family member	134	
Academic administration		
Qualification type	1	
High school academics		
Grade 11 Nov average	2	
Grade 12 June average	9	
Grade 12 Nov average	1	

7.5 Discussion of the secondary NGO sample data

The secondary industry partner also functions as an NGO and similarly facilitates bursaries on behalf of trusts, foundations, and private and corporate donors.

The secondary industry partner provided a data sample originally containing 407 science or engineering students who started their tertiary studies between 2010 and 2013. The variable names of this data sample, to be used in Case Study C, may be seen in Table 7.18. Each of the fields in this table are expanded upon in this section.

7.5.1 Personal information

The fields of *Gender* and *Race* indicate the gender and race of the students, respectively. The available categories within each of the two variables are shown in Tables 7.19–7.20, respectively.

7.5.2 Academic administration information

The field of *High school institution* indicates at which high school a student completed his or her Grade 12 studies. Within the sample, 71 different high schools were identified.

The *Tertiary institution* field represents the name of the tertiary institution that a student attended. The six tertiary institutions listed in Table 7.21 were identified in the sample.

The field *Academic status* indicates the students' final recorded academic situations. A student is assigned one of two main academic statuses: either *Completed* or *Not completed*. *Completed* indicates that the student has successfully completed his or her tertiary degree. *Not completed* indicates that a student is either still pursuing his or her qualification or has withdrawn from their tertiary studies. The two main statuses of *Completed* and *Not completed*, along with the possible reasons for not having completed, are shown in Table 7.22. *Academic exclusion* implies that the student was prohibited by the tertiary institution from continuing with his or her studies due to their poor academic record. The label *Declined* means the student declined a bursary offer. *Dropped out* implies the students withdrew from their tertiary studies, while *Still studying* means that the student is still in the process of pursuing his or her tertiary qualification (*i.e.* he or she has not graduated yet, but has also not withdrawn yet).

The *Study field* of a student indicates the academic field or discipline in which the student engaged their tertiary studies. The sample at hand only contains students who either studied *Engineering* or *Science*.

TABLE 7.17: *Number of missing entries for each independent variable in the sample data of Case Study B.*

	Data field	Number of entries missing from field
Personal information	Gender	0
	Race	3
	Family income	165
	Number of members in family	19
	Family income per member	150
Geographical	Home/Source region	95
	Tertiary study region	33
Academic administration	High school institution	57
	Tertiary institution	87
	Study field	0
	Qualification type	15
	Qualification	0
High school academics	Grade 11 Nov Average	760
	Grade 12 June Average	776
	Grade 12 Nov Average	158
High school subjects	Accounting	834
	Afrikaans (1st)	997
	Afrikaans (2nd)	931
	Agricultural science	986
	Business economics	823
	Economics	892
	English (2nd)	199
	Geography	725
	History	990
	IsiXhosa (1st)	949
	IsiZulu (1st)	826
	Life sciences	478
	Mathematics	406
	Mathematics literacy	969
	Physical science	557
	Setswana (1st)	977

The *Qualification* of a student refers to the exact degree programme for which he or she was enrolled (that is, the qualification they attempted to obtain or did, in fact, obtain, such as BEng Civil Engineering). There are 52 different qualifications listed in the sample.

The *Specialisation* field indicates the area within the study field in which the student specialises, such as Geology or Mining. Twenty four different specialisations are listed in the sample.

7.5.3 High school academic information

The three fields *Grade 12 Mathematics*, *Grade 12 Physical science*, and *Grade 12 English* contain the marks obtained by the students for Mathematics, Physical science, and English, respectively, in their Grade 12 year of high school.

Admission point score (APS) is calculated by each tertiary institution for their applicants based on their Grade 12 marks. Each tertiary institution calculates its APS score differently (for example, some include *Life orientation* while other exclude it from the calculation).

TABLE 7.18: *Fields in the sample data used for Case Study C.*

	Number	Data field	Data type
Personal information	1	Gender	Dichotomous
	2	Race	Multichotomous
Academic administration	3	High school institution	Multichotomous
	4	Tertiary institution	Multichotomous
	5	Academic status	Ordinal
	6	Study field	Multichotomous
	7	Qualification	Multichotomous
	8	Specialisation	Dichotomous
High school academics	9	Grade 12 Mathematics	Discrete
	10	Grade 12 Physical science	Discrete
	11	Grade 12 English	Discrete
	12	Admission point score (APS)	Discrete

TABLE 7.19: *Gender options applicable to Case Study C.*

Gender options	
1	Male
2	Female

TABLE 7.20: *Race options applicable to Case Study C.*

Race options	
1	Black
2	Coloured
3	Indian
4	White

TABLE 7.21: *Tertiary institutions supported by the secondary industry partner and applicable to Case Study C.*

Tertiary institutions supported			
1	Stellenbosch University	4	University of KwaZulu-Natal
2	University of Cape Town	5	University of Pretoria
3	University of Johannesburg	6	University of the Witwatersrand

TABLE 7.22: *Academic statuses applicable to Case Study C.*

Academic status			
1	Completed		
2	Not completed		
2.1	Academic exclusion	2.3	Dropped out
2.2	Declined	2.4	Still studying

7.6 Case Study C aim and variable selection

The data sample for Case Study C was described in §7.5. The data from this sample will also be analysed with the intent of trying to identify general trends and predictive variables.

7.6.1 Dependent and independent variable selection for Case Study C

For this case study, the dependent variable has been selected as the *Academic status* field, the categories of which were presented in Table 7.22. Those students with an *Academic status* of *Completed* were considered academically successful and were therefore assigned the value 1 for the dependent variable. The remainder of the students had the *Academic status* label of *Not completed* — for various reasons. Due to the nature of the study, only students who had either completed or withdrawn from their tertiary studies were considered, and no student currently still pursuing his or her tertiary studies was considered. For this reason, students with the *Not completed* label of *Still studying* were not considered. Those students with the *Not completed* label of *Declined* were also not considered due to the fact that they had left the bursary programme for reasons other than poor academics. One may also not assume that they did or did not go on to graduate successfully as no data are available in this respect. Only those remaining students with the *Not completed* label of either *Academic exclusion* or *Dropped out* were considered as having failed academically and were thus assigned the value of 0 for the dependent variable. The remaining eleven variables in Table 7.18 are considered candidate independent variables.

7.6.2 Data preparation and data cleaning of Case Study C

The relevant data set was obtained from the secondary industry partner in a tabular format within `Excel.xlsx` formatted files. The data files were well maintained and either did not require any integration or else only required minimal extraction as the files were saved over in a `Excel.csv` format so as to be used for input in the DSS concept demonstrator of Chapter 6.

The data set provided by the secondary industry partner originally contained a sample of 407 students. First, the dependent variable (*Academic status*) was assessed. Of the 407 students, 309 had an academic status which was not one of the desired types required in order to be able to be labelled either successful or unsuccessful. In addition, two of the students who had the *Not completed* label of *Dropped out* were excluded due to poor academics that were a result of poor health and anxiety disorder issues. Because this is not being a direct reflection of their academic ability, they were removed from the sample. Only 96 students remained in the sample after these removals.

Erroneous data records were also sought and corrected within the data set. Data entries were found to be erroneous because they contained valid error values and/or invalid error values. Each of these types of erroneous entries may be corrected by having the specific entry removed, replaced, or corrected. The number of cases of each type of erroneous data entry was identified within the sample and the corrective strategy chosen for each type is shown in Table 7.23.

The number of each type of erroneous entry observed for each applicable independent variable is shown in Table 7.24. Note that in this case study the "Remove" option again does not relate to removing the entire observation, but rather only the data in that specific field of the particular student. Thus none of the observations were removed and the sample still contained 96 students at the end of the cleaning process.

TABLE 7.23: *Identification of valid and invalid error values and applicable corrective strategies in the sample data for Case Study C.*

	Number	Chosen corrective strategy:		
	of cases	Removed	Replaced	Corrected
Valid error values	9			✓(9)
Invalid error values	1	✓(1)		

The erroneous entries of outliers and missing values are not considered at this stage of the case study data preparation for the same reasons as was explained in §7.4.2. The complete list of independent variables available to the analysts during this case study, and the number of entries missing form each, may be seen in Table 7.25.

Table 7.24: *Numbers of independent variables corrected due to erroneous values present in the sample data of Case Study C.*

	Valid error values	Invalid error values
Personal information		
Gender	7	
Academic administration		
Tertiary institution	1	
Qualification		1
Study field	1	

Table 7.25: *Number of missing entries for each independent variable of the sample data for Case Study C.*

	Data field	Number of entries missing from field
Personal information	Gender	0
	Race	1
Academic administration	High school institution	17
	Tertiary institution	5
	Academic status	0
	Study field	0
	Qualification	0
	Specialisation	1
High school academics	Grade 12 Mathematics	0
	Grade 12 Physical science	0
	Grade 12 English	0
	Admission point score (APS)	3

7.7 Case Study D aim

In the final case study, Case Study D, the capability and performance of the DSS will be assessed using data from both industry partners. Using different scenarios (*i.e.* selecting different independent variables and partitioning the data into learning and validation sets using randomly generated seed values), the performance of the weighted prediction of the DSS of Chapter 5 will be compared against those of the other base models. The scenarios to be considered will be chosen as those which showed the most promise during the analysis of the samples used in Case Studies B and C.

7.8 Chapter summary

In this chapter, data obtained from two industry partners were described along with the aim of each of four case studies to be performed in respect of these data in the following chapter.

In the first section, the target population for the study was identified. A discussion took place in the next section on the variables of the data sets provided by each of the two industry partners. This discussion included how each of the data samples were prepared for analysis. After each discussion of the sample data, the aim and variable selection of the case study to be performed in respect of the data samples were discussed. In total, the aims of four case studies were described, the first three being concerned with assessing variable importance and the fourth one with assessing the performance of the DSS proposed in Chapter 5 and conceptualised in Chapter 6.

CHAPTER 8

Case Study analyses

Contents

8.1 Case Study A 163
8.2 Roadmap for assessing variable importance 171
8.3 Case Study B 174
8.4 Case Study C 219
8.5 Case Study D 238
8.6 Chapter summary 242

The purpose of this chapter is to perform the four case studies introduced in Chapter 7. Case Study A, introduced in §7.3, is conducted first in §8.1, the purpose of which is to investigate the importance of three specific variables currently used in the bursary application process of the main industry partner. Based on the results of the case study, recommendations to the industry partner are formulated at the end of §8.1.

In the next section, a nine-step roadmap is presented for the general analysis of variable importance of all variables of Case Studies B and C present in the respective samples. Step seven of the roadmap involves using logistic regression models, random forest variable importance scores, CART plots, and frequency bar plots to investigate various models consisting of variables of the data.

Case Study B, introduced in §7.4 and conducted in §8.3, involves the investigation of a sample of 1101 observations for possible trends and important variables. Case Study C, conducted in §8.4, similarly involves using a smaller sample of 96 observations, and also utilises the nine-step roadmap of §8.2 to analyse its data.

The final case study, Case Study D, is conducted in §8.5. It involves using the sample data of Case Studies B and C to assess the capabilities of the DSS framework proposed in Chapter 5. The chapter concludes with a brief summary of the work conducted therein.

8.1 Case Study A

The data for and aim of Case Study A were presented in §7.2 and §7.3, respectively. Based on a data sample provided by the main industry partner, the variable importance of three score variables currently used in their bursary application process is assessed along with the importance of other variables typically also considered.

8.1.1 Reference class selection and assumption validation

The first step of the analysis is to select reference classes for any qualitative independent variables and assess the statistical assumptions of the sample data and models to be used.

For the only qualitative independent variable, *Qualification type*, its reference class is chosen as the category *Extended degree* as this is the category with the smallest proportion of observations having the desired outcome label.

The sample size for Case Study A consists of 109 observations. According to (3.78), the minimum required sample size for a logistic regression model containing seven independent variables, one of which has three categories, is in the order of 189 observations. Although the current sample size falls short of this size, the analysis will continue, but the small sample size will again be noted during feedback of the results. With respect to the minimum sample size required for random forests, it is recognised that the method of random forests is able to deal with reasonably small sample sizes. Although 109 observations is not ideal, the variable importance results obtained from random forests for a sample of 109 observations may be considered reasonably stable.

As discussed in §7.3, no outliers or influential observations were found in the sample and thus their potential removal was not necessary.

The seven independent variables to be used in this model are evaluated according to the assumptions of independence of residuals, multicollinearity, and linearity of the logit. No violations are found for the independence of residuals and multicollinearity assumptions, but the linearity of the logit assumption is violated for the quantitative variables of *Grade 11 November average* and *Interview score*. The result of this violation is that the statistical significance, as indicated by the logistic regression model, may underestimate the importance of these two variables. As the model will also be evaluated using random forest variable importance scores, these two variables will be left in and their results compared between the logistic regression model and random forests scores.

8.1.2 Evaluating the overall logistic regression model fit

The next step is to evaluate the overall fit of the logistic regression model. This is achieved by comparing the full logistic regression model against a null model. The result is a log likelihood p-value of 0.000203. This indicates that there exists at least one statistically significant independent variable (according to the p-value) within the logistic regression model and thus it encourages further investigation.

8.1.3 Evaluating individual independent variables

Due to the significant p-value in the above analysis, the analysis of the logistic regression model is continued by employing techniques that allow for the evaluation of the individual independent variables in the following three subsections.

Evaluating the p-values, odds ratios, and CIs

First, the p-values, coefficients, $\hat{\beta}$-CIs, and OR CI are determined for each of the independent variables in the logistic regression model. These values are provided in Table 8.1.

Based on only the p-values, it would appear that only the variable *Grade 11 November average* and the category *National diploma* of the variable *Qualification type* are of interest.

Grade 11 November average has a small p-value of less than one percent and large $\hat{\beta}$-coefficient which translates to a very large OR. Based on this OR, it may be stated that the odds of an individual being successful at tertiary level who pursues a *National diploma* over the odds of an individual being successful at tertiary level who pursues an *Extended degree* (the reference class) is approximately 40, while keeping the other independent (control) variables constant. The CI width of this category is, however, very large, suggesting a lack of precision present and so this large OR may well be overestimated. It is, however, not possible to make any pronouncement about whether the two categories of *National diploma* and *Degree* are significantly different in respect of predicting tertiary success of students.

The second most interesting variable is *Grade 11 November average*, which has a p-value of less than one percent. Although this p-value is relatively small, its $\hat{\beta}$-CI width is very small in comparison to that of *National diploma* and thus indicates good precision. The CI of this variable does contain zero in this case, which is due to the usage of a 95% CI and a p-value greater than 5%. If a 90% CI were to be used, this containment would not occur. Potentially the most interesting result in the table is that the $\hat{\beta}$-coefficient of *Grade 11 November average* is negative, although only slightly.

Despite the above observations, the reader should keep in mind that due to the violation of the linearity of the logit assumption, the *Grade 11 November average* variable may well have been allocated a larger p-value than it should have. The importance of the *Grade 11 November average* according to the logistic regression model is therefore compared with that of its random forest variable importance score in the next subsection.

TABLE 8.1: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for the logistic regression model of Case Study A.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.530478	2.064	−4.385	8.512	12.897	7.876	0.012	4 975.467	4 975.454
Form score	0.600009	−0.082	−0.388	0.224	0.612	0.921	0.679	1.251	0.572
Interview score	0.371922	−0.087	−0.278	0.104	0.383	0.917	0.757	1.110	0.353
Recommendation	0.262779	−0.465	−1.277	0.348	1.626	0.628	0.279	1.417	1.138
Grade 12 Nov	0.154482	0.060	−0.023	0.142	0.165	1.062	0.978	1.153	0.175
Grade 12 June	0.981213	0.001	−0.050	0.051	0.102	1.001	0.951	1.053	0.102
Grade 11 Nov	0.098817	−0.059	−0.129	0.011	0.140	0.943	0.879	1.011	0.132
Nat diploma	0.004110	3.673	1.164	6.182	5.018	39.386	3.204	484.161	480.957
Degree	0.211831	1.470	−0.838	3.778	4.616	4.350	0.433	43.736	43.303

Considering the random forest variable importance

Secondly, variable importance measures are calculated for each independent variable by an implementation of the method of random forests, described in §3.7.1. A model containing the same variables as were used for the logistic regression model above is again used in conjunction with random forests now. For the remainder of Case Study A the term ‘model’ refers to this model.

Unlike logistic regression, the method of random forests is apparently able to fit models that contain multicollinearity and be relatively unaffected by its presence. This allows for the construction of two variable importance score plots, shown in Figures 8.1–8.2, drawn up for the same model. The first plot contains one score for each of the independent variables. In the second plot, however, the qualitative variable *Qualification type* is converted into three dummy variables, one for each of its categories.

By investigating Figure 8.1 it should be clear that *Qualification type* is identified as the most important variable, within which the categories *Degree* and *National diploma* are the most important according to Figure 8.2. By considering both plots it would appear that the only important variable in respect of predicting tertiary success of students is *Grade 11 November average*. As in the logistic regression model, the remaining variables, and specifically the three NGO scores, carry disappointing importance scores.

It would also appear that the violation of the linearity of the logit assumptions does not negatively impact the p-values allocated to the *Grade 11 November average* and the *Interview score* variables during logistic regression due to the similarities of the p-value findings and the random forests scores.

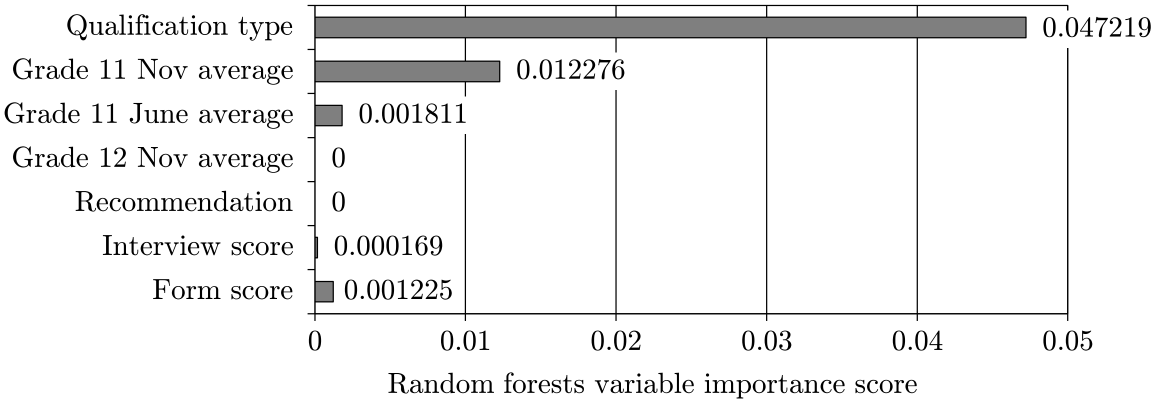


FIGURE 8.1: *Random forest variable importance scores for the model of Case Study A.*

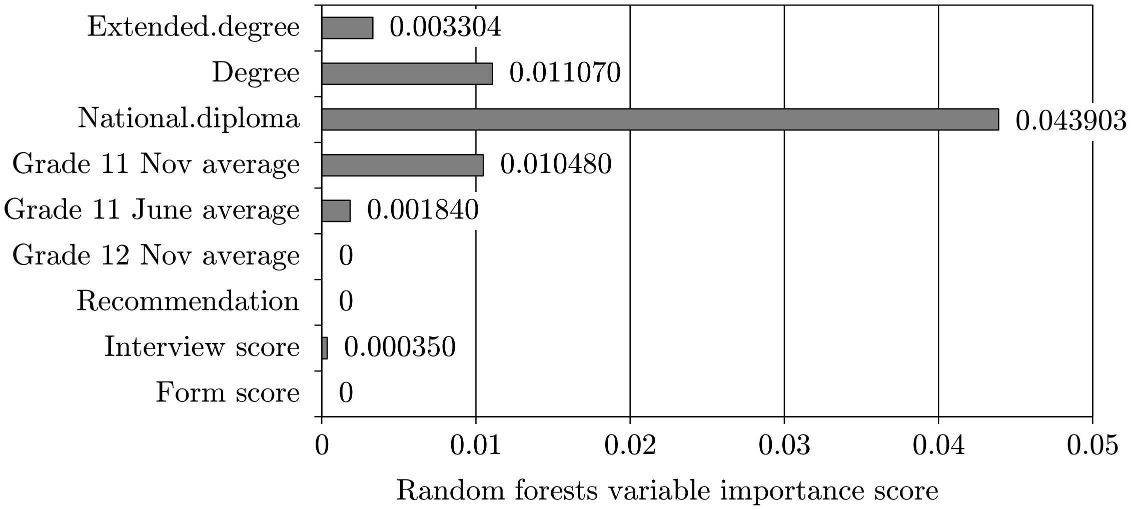
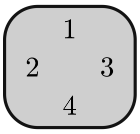


FIGURE 8.2: *Random forest variable importance scores for the model, including dummy variables, of Case Study A.*

Plotting the classification and regression tree

The third step in evaluating the individual independent variables is to construct a CART plot, shown in Figure 8.4. The four values for each node of the CART plot may be interpreted as shown in Figure 8.3. The background colour is red if the majority class has the label "0" and green if the majority class has the label "1".



1 Majority outcome label of node.
2 Proportion of observations in node that have the outcome label of “0”.
3 Proportion of observations in node that have the outcome label of “1”.
4 Percentage of sample that is included in node.

FIGURE 8.3: *Interpretation key for the CART plot nodes.*

The partitioning of the root node is unsurprising as the variable *Qualification type* is used. This is expected, since it was found to be the variable with the highest random forest variable importance score in Table 8.1. By investigating the further partitioning of the categories it would appear that those students pursuing a *National diploma* are far more likely to succeed at tertiary level than those pursuing a *Degree* or *Extended degree*.

The second partition is also expected as the second most important variable identified was *Grade 11 November average*. It is noted that a peculiar trend is present in which a higher *Grade 11 November average* for a student seems to lead to a higher probability of him or her being unsuccessful at tertiary level.

The final partition, based on the variable *Interview score*, is also counter-intuitive as it seems to indicate that those students who were assigned a higher *Interview Score* are more likely to withdraw from their studies.

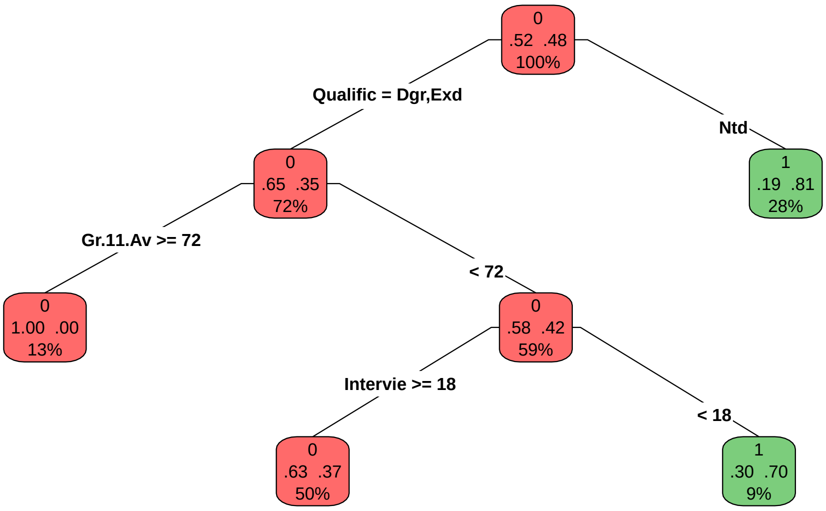


FIGURE 8.4: *CART plot for the model of Case Study A.*

8.1.4 Discriminant ability of the independent variables

The model’s prediction ability is next evaluated by having it trained with a bootstrapped subsample of the 109 observations, with replacement, and tested against the remaining OOB observations for 500 iterations. The average predictive accuracy of the weighted prediction produced by the DSS for the model is 71.11%.

8.1.5 Further analysis of the two most important independent variables

It is possible to analyse the two variables of *Qualification type* and *Grade 11 November average* further by plotting the percentage of students who achieved academic success pursuing one of the three degree types, and comparing this percentage to their Grade 11 November average marks. These two plots are shown in Figures 8.5–8.6.

In Figure 8.5, a clear distinction between the three categories of *Qualification type* is visible. This would indicate that the variable is good at distinguishing those students who will be successful from those who will not and thus agrees with the previous results that this variable exhibits predictive power in terms of academic success at tertiary level.

A visual examination of Figure 8.6, however, produces very interesting findings. After excluding the two outer bins containing only four observations each, a clear negative relation trend is visible, namely that as the Grade 11 November average mark of the applicant decreases, his or her tertiary success rate increases. Although this variable has been identified as important in the previous steps of the analysis, the fact that the relationship is negative is fascinating and rather counter-intuitive. This, of course, agrees with the negative $\hat{\beta}$-coefficient assigned to the *Grade 11 November average* in Table 8.1. Possible reasons for this are discussed in the next subsection.

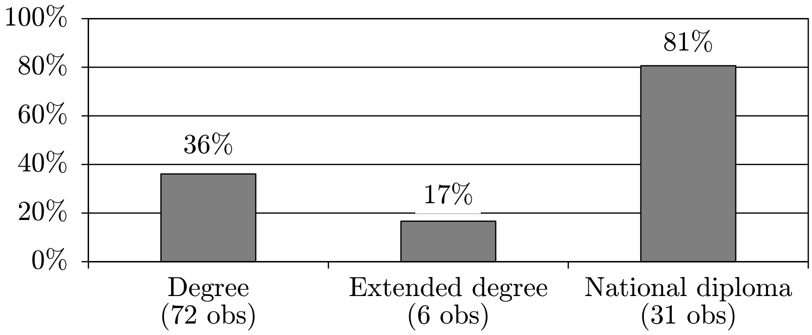


FIGURE 8.5: *Percentages of students who graduated successfully per qualification type in the sample of Case Study A.*

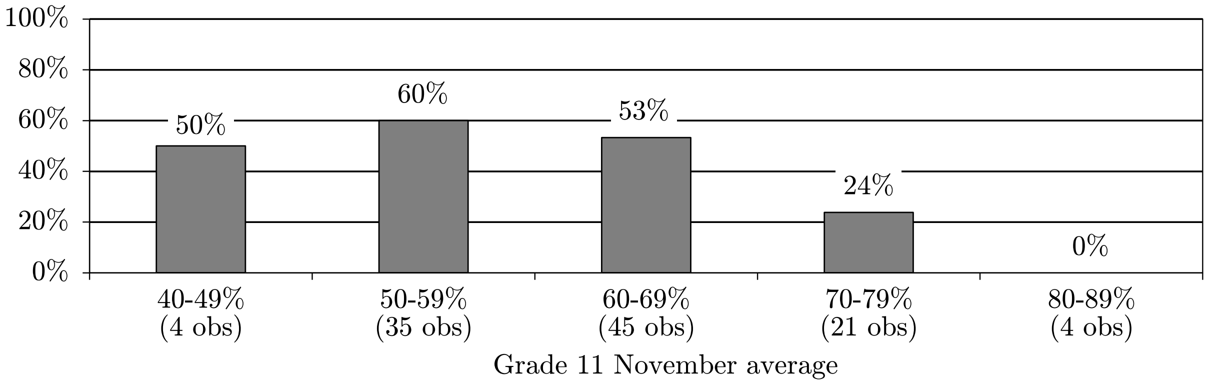


FIGURE 8.6: *Percentages of students who graduated successfully as a function of Grade 11 November averages in the sample of Case Study A.*

8.1.6 Results summary and recommendations

A concise summary of the above analysis is that *Grade 11 November average* and *Qualification type* are the most important variables in terms of possibly predicting successful tertiary graduation, while the predictive power of the three NGO score variables would seem to be comparatively insignificant.

A further discussion follows on the findings of the analysis and recommendations to the main industry partner are formulated based on the analysis findings, specifically focussing on the importance of the three NGO score variables, since this was the original aim of Case Study A.

Qualification type

The importance of *Qualification type* is a slight surprise. As shown in Figure 8.5, however, a possible reason for the predictive power of this variable is the fluctuation present between the categories as they relate to the academic success percentage of students. That is, the percentage of students who achieved academic success based on their qualification type fluctuates greatly based on whether a student enrolled for a degree, an extended degree, or a national diploma. It should be acknowledged, however, that although these variables are indirectly related to a student's academic ability, they are much more representative of the difficulty of students' tertiary degree or diploma choices, rather than of students' academic abilities. This is, to some extent, expected as it is generally accepted that degree qualifications are more challenging to obtain than national diplomas. It is difficult to make general assumptions about the category *Extended degree*, but it might be a qualification often attempted by students who only just qualified for admission to such a programme, in which case the low success rate associated with an extended degree programme is understandable.

It is advised, therefore, that the other variables more directly related to students' academic abilities should still be collected. Also, although no students should be barred from entering a course for which they have been accepted, it is advised that the industry partner should be more strict in respect of providing funding to students who are entering study fields, tertiary institutions or applying for degree types which have been shown to be more challenging for the industry partner's bursary holders. This should be done over and above the criteria already in place, as a student would automatically rank well above the regular applicants if he or she has been accepted into a more challenging study field, tertiary institution or degree type.

High school marks

The predictive power of the high school average mark results were unexpected. The importance of *Grade 12 November average* was disappointing, but it is believed that relating specific high school subjects to specific qualifications and degrees may well produce fruitful results. This could unfortunately not be done with the data sample of this case study, because it was too small. This line of investigation will, however, be pursued in Case Study B, for which the data sample is much larger.

The surprising result was the predictive power of *Grade 11 November average*. Not only did it rank as the only most significant variable of the high school marks, but as explained in §8.1.3, it would seem that a lower *Grade 11 November average* indicates a higher probability of tertiary academic success. This certainly does not mean that the industry partner should seek out and provide bursaries to those students with the lowest *Grade 11 November average* marks. There is a possibly plausible explanation for the finding. Of the three high school examination averages investigated, only the *Grade 12 November average* is the result of an examination set nationally and independently by the *Independent Examination Board*. Grade 11 November examinations, on the other hand, are set independently by each school. This, of course, results in large fluctuations of the standard of those examinations. It has been speculated that in many schools the teachers either set easier examinations or slightly adjust the marks of students upwards so as to allow them to gain entrance into the field of study of their choice. Whether or not these speculations carry merit, the fact is that students with higher Grade 11 examination marks are more likely to be provisionally accepted by tertiary institutions with higher standards or more challenging degree types. Once provisionally accepted by a tertiary institution, it is only required of those students to obtain the minimum required marks in their Grade 12 November examinations. As illustrated in Figure 8.5, the varying difficulties of qualification types would then come into play.

The three NGO scores

With respect to the three NGO score variables, it would seem that none of these variables offers any predictive assistance in respect of identifying who will be successful at tertiary level. Despite this finding, it may be possible simply to improve the collection of these scores so as to render them more adequate for predicting tertiary success. The following is a list of recommendations to the main industry partner for possibly improving the collection of these scores:

1. The number of categories of the recommendation score may be increased to (say) ten so as to create a wider spread of the scores.
2. The categories of the recommendation score, listed in Table 7.10, may be described in more detail so as to assist the recruiters in assigning scores more consistently. Currently two of the categories are associated with the same description.
3. In order to avoid personal bias, recruiters should only be allowed to conduct interviews and assign recommendation scores in areas outside their local communities. The recruiters should, however, obviously still be allowed to promote the bursary programme in their local communities and possibly assist with the completion of the initial bursary application forms there.
4. In order to avoid potential inconsistency during the interview process, the option of using fewer recruiters who would then cover more areas should be considered. Each additional recruiter has his or her own interpretation of the interview and recommendation scores. Another option, which is currently being implemented, is to continue to use more than one recruiter per interview.
5. As human error is a common occurrence during the process of data entry, it is advised that the scores entered by an individual into the system from the various forms be audited to some degree, even if just in a small way.
6. Emphasis should be placed on good data collection, as has only recently become standard practice at the main industry partner. This is an important step towards the facilitation of future statistical studies. If a data record is missing data on a single variable being investigated, the entire record of that student must be removed from the sample in a statistical analysis involving that variable!

As should be evident from the thorough data cleaning process described in §7.3.3, a proper attempt was made at preparing the data set adequately for the preceding analysis. Despite this attempt, it should be noted that the findings of this study are potentially hindered by the small size of the sample used in the sense that the results are reported with less confidence than would have been the case with a larger sample.

An analysis of the same industry partner's data follows in Case Study B later in this chapter, but using a larger sample. It will thus be possible to examine whether or not the interesting findings of this case study, namely that *Qualification type* is the most important predictive variable of the tertiary success of a student and that high school marks are negatively correlated to tertiary success, generalise to a larger data set.

8.2 Roadmap for assessing variable importance

A variable importance assessment roadmap is presented in this section which is to be utilised in Case Studies B and C later in this chapter. The term 'model' refers to a specific combination of a dependent variable and one or more independent variables in the remainder of this section. Depending on the context, the model may be a logistic regression model, a CART model, or a random forests model.

Numerous studies found in the literature consider many (hundreds or even thousands of) independent variables. Some of these studies often utilise a method such as stepwise logistic regression, which is sightly disputed, to reduce a model initially containing all the independent variables down to a model containing far fewer, but more statistically significant, variables [11, p. 1138], [107, p. 377], [193, p. 198], [215].

Due to the many missing values in the data samples of Case Study B, a 'clean' approach of assessing all independent variables together in a single model cannot be adopted. For example, two of the independent variables to be used in Case Study B are *Grade 12 Mathematics* and *Grade 12 Mathematics literacy*. If both are selected in a model, no observations will be left in the sample as no single student takes both these subjects.

Although a model exists that can fit all the variables for the sample data of Case Study C, an automated variable selection algorithm will not be employed due to the disputes surrounding its use. In addition, the number of variables available for Case Study C is not so large that it necessarily justifies the use of removing many of the variables.

Due to the peculiar make-up of the two data samples and the above reasons, the followings less than ideal nine-step roadmap for conducting Case Studies B and C is adopted as it has become evident from prior analyses of the data that little to no analysis can be performed if more traditional statistical approaches are followed and sample size requirements strictly adhered to. So as to provide the industry with adequate feedback based on the limited amount of information that is actually available, the conclusion was drawn that it would be better to provide the industry partners with the results of analyses of samples that are less recommended by the literature, but nevertheless to provide some feedback with caution due to small samples, as opposed to no feedback at all. Data are simply not abundant enough in two of the case studies so as to rule out the use of classification models due to a lack of variable values. Instead of ignoring certain models completely, the sample sizes will be noted and reported together with analysis results and subsequent recommendations so as to caution the user to consider the results in this light. It will also be kept in mind that p-values tend to decrease with increased sample size, and *vice versa*. This, in fact, will to some extent lend even more support to small p-values attached to a small sample. The exclusion of classification models due to the other assumption violations, such as those of multicollinearity, will be applied more strictly.

The following nine steps of the roadmap will only be performed once data preparation (gathering, integration, extraction, and cleaning) has taken place:

1. *Select a single dependent variable.* Identify a single, binary outcome dependent variable from the variables listed in the data.

2. *Identify all remaining individual independent variables and select reference classes.* Identify all remaining independent variables from the sample that are viable options for independent variables. The variables should, for example, not involve any information that can only be collected after a student has started their tertiary studies, such as their first semester tertiary marks. Moreover, the high school institution that a student attended is also not suitable as an independent variable, as there is generally a large number of schools listed in proportion to the number of students.

   For each of the qualitative independent variables, a reference class should be identified. The use of dummy variables and reference classes for qualitative independent variables were discussed in §3.3. It was noted that the reference classes may be selected as extreme cases as this simplifies the interpretation of the results. In the case that the number of observations in the reference class is relatively large it is an added bonus. In order to promote consistency, the category containing the lowest proportion of desired outcomes will always be selected in this thesis as the reference class.

3. *Find largest combinations of independent variables that have a sufficient number of observations and do not exhibit extreme multicollinearity.* Attempt to create the largest possible combinations of variables until all the independent variables are included in at least one combination. Ideally one will be able to include all variables in a single model, but due to missing values in some of the variables and multicollinearity between other variables this will not always be possible. The variable combinations should not contain variables exhibiting extreme multicollinearity and the minimum sample size assumption should preferably not be violated for reasons explained above. A single combination should, however, contain no fewer than approximately 90 observations, irrespective of the number of independent variables. The sample sizes of the combinations should ideally be assessed using stable logistic regression guideline of (3.78).

   In addition to attempting to create the largest possible logistic regression models, attempts should also be made to build models for each independent variable separately. If there exist m independent variables, after this step there should exist at the least $m+1$ models. For cases in which strong multicollinearity is present between two variables in a combination containing, say five, independent variables, the combination may be 'duplicated.' The term 'duplicated' here refers to the practice of creating two of the above combinations in which multicollinearity is present by removing the first variable associated with the multicollinearity from the one new combination and removing the second variable associated with the multicollinearity from the other new combination.

4. *Identify outliers and influential values for the remaining combinations and consider removal of the identified observations.* The remaining models should be assessed for outliers and influential values. A discussion on outliers and influential values, how to identify them, and the advantages and disadvantages of removing or keeping them, may be found in §2.2.4 and §3.7.2.

   Three options are available. The models for which outliers and influential values were identified may be removed, the models may be kept, but the outliers and influential values may be removed, or both the outliers and influential values and the models may be kept.

   The sample size of combinations should also be taken into consideration when deliberating whether or not to remove the identified outliers and influential values, as their removal might drastically change the sample size.

   In the case that outliers or influential observations are removed, the multicollinearity of those combinations should be evaluated again as it might have increased due to the removal of the outliers or influential observations.

5. *Consider the other assumptions.* For the remaining models, the other underlying assumptions should be validated and removal of the models for which these assumptions are not satisfied should be considered. These other assumptions include testing for the independence of residuals (see §3.10.2), testing for linearity of the logit (see §3.10.3), and ensuring that minimum sample sizes are preferably adhered to. Minimum sample sizes are relevant for logistic regression (see §3.10.3), random forests (see §3.10.5), and CART (see §3.10.4).

   Violation of a minimum sample size should not automatically result in the removal of that model, but the sample size results should be considered within the context of the case study. For some of the case studies there will not be many available data entries, but performing the statistical tests on those models should still be considered as there simply may not exist enough data to satisfy the minimum sample size requirements set out by the literature. After all, a model with little data whose results are presented with a warning message due to few entries is still better than no model results. Thus, the minimum sample results should be reported appropriately in conjunction with the results of the last step of this roadmap.

6. *Evaluate the overall model fit using the method of log likelihood and create new combinations of best variables.* For the remaining models, determine their overall fit using the notion of log likelihood as outlined in §3.5.1. Remove any model that does not produce a significant p-value (that is, a p-value which does not fall below the set α-threshold). For the remaining models it is assumed that at least one of the independent variables present in the model is significant.

7. *Evaluate the individual independent variables within the remaining models.* Assess the individual independent variables within the remaining models using the following three steps:

   (a) *Determine and evaluate the p-values, $\hat{\beta}$-coefficients, odds ratios, and their CIs.* For each individual independent variable within the remaining models, determine and evaluate the p-values, $\hat{\beta}$-coefficients, odds ratios, and their CIs, as was described in §3.5.2. The term *statistically significant* will be used to indicate when a specific variable or category obtains a p-value below the set α of 5%, thus indicating adequate evidence against the H_0 so as to reject it. The H_0 for Case Studies B and C state that $\hat{\beta}$-coefficients of the variable or category is equal to zero, *i.e.* they do not have any predictive ability.

   (b) *Determine and evaluate the random forest variable importance scores.* For the models that exhibit significant p-values according to the log likelihood test in Step 6, determine their random forest variable importance scores, as outlined in §3.7.1. Combinations that were duplicated in Step 3 due to the presence of multicollinearity may be combined for the random forest variable importance score assessment since they are apparently unaffected by multicollinearity.

   (c) *Determine and evaluate the CART plot associated with each model.* Using the method of CART (as described in §3.6), plot decision trees and evaluate them.

   (d) *Construct frequency bar plots of the variables of interest.*

8. *Perform bootstrapping predictive accuracy tests on promising models.* For those models that show promise in being able to identify tertiary successful students, based on the evaluations of Step 7, perform bootstrap predictive accuracy assessments and evaluate the results in the manner discussed in §3.5.3.

9. *Report back results.* Interpret and report back all findings in an understandable manner.

8.3 Case Study B

The data for and aim of this second case study were presented in §7.2 and §7.4, respectively. The roadmap for assessing variable importance, outlined in §8.2, is now applied in a stepwise manner to these data. The nine subsections of this section are directly associated with the nine steps of the roadmap.

TABLE 8.2: *Independent variables and reference classes for the sample data of Case Study B.*

	Field number	Data field name	Reference class	Proportion of reference class outcome labelled 0
Personal information	1	Gender	Male	52%
	2	Race	Coloured	56%
	3	Family income per member	N.A.	N.A.
Academic administration	4	Study region	Gauteng	47%
	5	Tertiary institution	University of Pretoria (UP)	31%
	6	Study field	Engineering	40%
	7	Qualification type	Extended degree	45%
High school academics	8	Grade 11 Nov Average	N.A.	N.A.
	9	Grade 12 June Average	N.A.	N.A.
	10	Grade 12 Nov Average	N.A.	N.A.
	11	Accounting	N.A.	N.A.
	12	Afrikaans (1st)	N.A.	N.A.
	13	Afrikaans (2nd)	N.A.	N.A.
	14	Agricultural Science	N.A.	N.A.
	15	Business economics	N.A.	N.A.
	16	Economics	N.A.	N.A.
	17	English (2nd)	N.A.	N.A.
	18	Geography	N.A.	N.A.
	19	History	N.A.	N.A.
	20	IsiXhosa (1st)	N.A.	N.A.
	21	IsiZulu (1st)	N.A.	N.A.
	22	Life sciences	N.A.	N.A.
	23	Mathematics	N.A.	N.A.
	24	Mathematics literacy	N.A.	N.A.
	25	Physical science	N.A.	N.A.
	26	Setswana (1st)	N.A.	N.A.

8.3.1 Selection of a single dependent variable

The variable *Academic status* was identified as the dependent variable, as described in §7.4.1.

8.3.2 Identification of independent variables and reference classes

The thirty one possible independent variables for the sample of 1101 observations were shown in Table 7.17. Some of these variables were, however, excluded from consideration for the following reasons. The variable *Family income* does provide valuable information, but even more so when considered alongside the *Number of members in family*. In the same fashion, the *Number of members in family* variable is even more informative when considered in conjunction with *Family income*. For these reasons, the variable *Family income per member* is considered as it is produced by the combination of the variables *Family income* and *Number of members in family*. That is, only the *Family income per member* variable is considered, while the variables *Family income* and *Number of members in family* are not considered.

The *Source region* variable, presented in §7.2.2, contains the province and region, or city, within that province, from which the students originate. In the sample of 1101 observations, 114 different *Source regions* are listed. Due to the large ratio of *Source regions* to observations, this variable is not considered.

Similarly, the *High school institution* field contains 492 different high schools and the *Qualification* field contains 194 different categories. Again, these large ratios make both prediction and variable assessment impractical for these two variables, and so neither is considered. After these exclusions, only twenty six independent variables remain, as listed in Table 8.2.

As explained in §8.2, qualitative variables are converted to multiple levels using the method of dummy variables. It was also explained that the reference class or baseline class for each categorical variable is selected as the category with the lowest proportion of desired outcome labels. The selection of reference classes for the remaining qualitative independent variables may also be seen in Table 8.2.

8.3.3 Identification of largest valid combinations of remaining variables

In the next step, the relationships between the independent variables are investigated so as to identify which of the relationships exhibit multicollinearity. Strong multicollinearity is unsurprisingly found between the variables *Tertiary institution* and *Study region*, as each tertiary institution is only found in one study region (province).

In addition, extreme multicollinearity is found to be present between *Tertiary institution* and *Grade 11 November average*, and between *Tertiary institution* and *Grade 12 June average*, but not between *Tertiary institution* and *Grade 12 November average*. The reason for this finding may be that many tertiary institutions preliminarily accept students based on their *Grade 11 November average* marks and their *Grade 12 June average* marks, and are then often only required to obtain the minimum requirement marks for the *Grade 12 November average*.

It was also noticed that although the multicollinearity score between *Tertiary institution* and *Qualification type* does not exceed the violation limit, it is still relatively high. A possible reason for this may be that nine of the tertiary institutions do not offer national diploma qualifications. Similarly, a medium level of multicollinearity is detected between *Tertiary institution*, *Qualification type*, and *Study field*. This is of course expected, but the combinations of independent variables are not partitioned further to accommodate these findings, as this will result in almost no large variable combinations being considered. These findings will, however, be considered when evaluating the results. The above are only some of the examples of variable combinations for which medium multicollinearity is detected. The nature of the sample data, with a large number of the variables displaying medium multicollinearity, has unfortunately made the complete avoidance of multicollinearity during the construction of useful combination unavoidable.

The most likely consequences of the medium levels of multicollinearity between the variables are that specific categories within the variables will be affected, since they are all qualitative variables. If the study was aimed at reporting back specifically which categories of variables are important, this would be a major problem, but in this study the aim is to investigate variables as a whole, and only broadly. The alternative is, of course, only considering models with single variables, but that would potentially hinder the discovery of interesting findings. The model combinations are constructed so that no model would contain any two variables found to exhibit extreme multicollinearity (exceeding a VIF of 10), but combinations containing medium levels of multicollinearity are, as explained above, unavoidable.

In addition to the presence of multicollinearity, the sample size of each combination is also considered in conjunction with constructed variable combinations, but combinations are not excluded due to violation of sample size as this would inhibit the construction of particularly interesting combinations. The combinations formed by combining the individual variables of Table 8.2 may be seen in Table 8.3.

TABLE 8.3: *Number of available and required observations for each variable combination of Case Study B.*

				Number of observations:	
Combination number	Fields in combination	Number of fields	Number of variables and categories	in each combination	required by (3.78)
1	1	1	2	1101	49
2	2	1	2	1098	49
3	3	1	1	951	25
4	4	1	7	1068	168
5	5	1	16	1014	370
6	6	1	12	1101	289
7	7	1	3	1086	73
8	8	1	1	341	21
9	9	1	1	325	22
10	10	1	1	943	25
11	11	1	1	267	32
12	12	1	1	104	23
13	13	1	1	170	34
14	14	1	1	115	23
15	15	1	1	278	30
16	16	1	1	209	33
17	17	1	1	902	24
18	18	1	1	376	22
19	19	1	1	111	25
20	20	1	1	152	25
21	21	1	1	275	31
22	22	1	1	623	23
23	23	1	1	695	22
24	24	1	1	132	29
25	25	1	1	544	21
26	26	1	1	124	23
27	1,2,3,4,6,7	6	27	908	643
28	1,2,3,5,6,7	6	36	863	858
29	1,2,3,4,6,7,8,9,10	9	30	254	715
30	1,2,3,5,6,7,10	7	37	761	881
31	8,9,10	3	3	294	72

The first twenty six combinations contain only single variables, so as to assess the individual performance of the twenty six variables listed in Table 8.2.

The next two combinations, 27 and 28, both contain the variables *Gender*, *Race*, *Family income per member*, *Study field*, and *Qualification type*, but only Combination 27 also contains *Study* *region*, while only Combination 28 also contains *Tertiary institution*. Combinations 27 and 28 are constructed in this manner since *Study region* and *Tertiary institution* are unsurprisingly singular.

Combination 29 is identical to Combination 27, and Combination 30 is identical to Combination 28, except that Combinations 29 and 30 also contain *Grade 12 November average*. Combination 29 furthermore contains *Grade 11 November average* and *Grade 12 June average*, but Combination 30 may not, since *Tertiary institution* is highly correlated with both. Combination 31 contains the variables *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*. The remainder of the high school subjects are to be considered for new combinations after the overall model fit is evaluated (using log likelihood p-values).

Of the thirty two combinations, only Combinations 29 and 30 do not satisfy the minimum sample size requirements for logistic regression. As explained in §8.2, however, these combinations are nevertheless considered, but the results thus produced are presented cautiously.

8.3.4 Identification of outliers and influential values

The number of outliers and influential observations identified for each of the thirty one combinations may be viewed in Table 8.4.

8.3.5 Validation of the remaining assumptions

Once the decision was made as to whether or not to delete the outliers and influential observations, the remaining three assumptions of independence of residuals, multicollinearity, and linearity of the logit were considered for each of the nine combinations, the results of which may also be seen in the last three columns of Table 8.4.

Combinations 3, 27, 28, and 30 were identified as containing a large number of outliers, which made up 10.7%, 17.7%, 18.5%, and 20.0% of their respective samples. Due to the fact that the author would like to analyse as much of the original data as possible and report results that accurately reflect real trends, the removal of such a large percentage of the data was deemed unwise, and thus the identified outliers were retained subjectively.

On the other end of the spectrum, Combinations 8, 9, 10, 11, 16, 17, 18, 22, and 25 were identified as containing very small numbers of outliers. As before, the author believes that by not deleting these small numbers of outliers, which should in actual fact not severely influence the results of the analysis, more reliable results will be obtained. This will also be beneficial for consistency with respect to the handling of outliers. If such small numbers of influential values were identified, their removal would have been much more likely.

It was a surprise that high levels of multicollinearity were not detected in Combinations 29 and 31 between *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*. Upon closer inspection it became clear that the marks obtained for the three variables fluctuate significantly for each student. No other cases of significant multicollinearity are detected.

8.3.6 Evaluation of the overall model fit and addition of new combinations

Each of the thirty one combination models are assessed in this section against a null model so as to obtain the log likelihood p-values shown in Table 8.5. Those combinations showing statistical significance at a 5%-level of significance are typeset in boldface. Based on the log likelihood p-values produced, a further three combinations are constructed, as shown in Table 8.6.

The three most promising individual high school subjects of *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science* are added to the three high school averages of

Table 8.4: *Number of outliers and influential observations identified, and assessment of the remaining assumptions, for each combination of Case Study B.*

			Actions taken:		Assumption of:		
Combination number	Num of outliers	Num of inf obs	Outliers deleted?	Inf obs deleted?	Independence of residuals	Multi-collinearity	Linearity of the logit
1	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
2	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
3	102	0	No	N.A.	Satisfied	N.A.	Satisfied
4	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
5	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
6	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
7	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
8	1	0	No	N.A.	Satisfied	N.A.	Satisfied
9	2	0	No	N.A.	Satisfied	N.A.	Satisfied
10	1	0	No	N.A.	Satisfied	N.A.	Satisfied
11	1	0	No	N.A.	Satisfied	N.A.	Satisfied
12	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
13	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
14	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
15	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
16	1	0	No	N.A.	Satisfied	N.A.	Satisfied
17	7	0	No	N.A.	Satisfied	N.A.	Satisfied
18	5	0	No	N.A.	Satisfied	N.A.	Satisfied
19	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
20	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
21	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
22	7	0	No	N.A.	Satisfied	N.A.	Satisfied
23	8	0	No	N.A.	Satisfied	N.A.	Satisfied
24	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
25	9	0	No	N.A.	Satisfied	N.A.	Satisfied
26	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
27	161	0	No	N.A.	Satisfied	Satisfied	Satisfied
28	160	0	No	N.A.	Satisfied	Satisfied	Satisfied
29	0	0	N.A.	N.A.	Satisfied	Satisfied	Satisfied
30	151	0	No	N.A.	Satisfied	Satisfied	Satisfied
31	0	0	N.A.	N.A.	Satisfied	Satisfied	Satisfied

Grade 11 November average, *Grade 12 June average*, and *Grade 12 November average* to form Combination 32. It is unfortunately not possible to form new combinations with the other promising high school subject of *Grade 12 Accounting* due to its small sample size and few overlapping entries between observations of other promising variables.

It also seems desirable to combine the three promising high school subject variables not only with the high school averages, but also with the other variables that produced small log likelihood p-values. Based on this, Combinations 33 and 34 were created containing variables 1, 3, 4, 5, 6, 7, 10, 17, 23, and 25. As before, variables 4 and 5 are each only present in one of the new combinations due to multicollinearity. Unlike Combination 32, these two new combinations do not contain *Grade 11 November average* and *Grade 12 June average* as their inclusion would result in an even smaller sample size. As with some of the other combinations, the sample size for Combinations 33 and 34 is not adequate, but as previously explained, analysis of these combinations will continue cautiously.

The assessment of the number of outliers and influential values, as well as the validity of the three other assumptions in the context of the new combinations are shown in Table 8.7. A few outliers are detected as being present in Combinations 33 and 34, but by the same logic as before they are not deleted.

TABLE 8.5: *Log likelihood p-values for each combination of Case Study B.*

Combination number	Log likelihood p-value	Combination number	Log likelihood p-value	Combination number	Log likelihood p-value
1	**0.000003**	12	0.578158	**23**	**0.000001**
2	0.706823	13	0.873780	24	0.874844
3	**0.006614**	14	0.720248	**25**	**0.000243**
4	**1.33×10^{-9}**	15	0.830143	26	0.606557
5	**5.86×10^{-21}**	16	0.220421	**27**	**2.35×10^{-25}**
6	**1.80×10^{-14}**	**17**	**0.028013**	**28**	**1.15×10^{-24}**
7	**2.28×10^{-14}**	18	0.556528	**29**	**3.47×10^{-13}**
8	**0.008681**	19	0.134165	**30**	**6.38×10^{-30}**
9	**0.119618**	20	0.957433	**31**	**0.016341**
10	**0.002543**	21	0.603928		
11	**0.019738**	22	0.615169		

TABLE 8.6: *Number of available and required observations for each additional combination of Case Study B.*

				Number of observations:	
Combination number	Fields in combination	Number of fields	Number of variables and categories	in each combination	required by (3.78)
32	8,9,10,17,23,25	6	6	183	143
33	1,3,4,6,7,10,17,23,25	9	29	430	691
34	1,3,5,6,7,10,17,23,25	9	38	417	905

TABLE 8.7: *Number of outliers and influential observations identified, and assessment of the remaining assumptions, for each additional combination of Case Study B.*

			Actions taken:		Assumption of:		
Combination number	Num of outliers	Num of inf obs	Outliers deleted?	Inf obs deleted?	Independence of residuals	Multi-collinearity	Linearity of the logit
32	0	0	No	N.A.	Satisfied	Satisfied	Satisfied
33	26	0	N.A.	N.A.	Satisfied	Satisfied	Satisfied
34	23	0	No	N.A.	Satisfied	Satisfied	Satisfied

Just as it was a surprise above that multicollinearity is not detected between *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*, it was again slightly unexpected that in Combination 32 there are also no severe multicollinearity violations present between the high school averages and high school subject marks. A possible reason for this might be that although *Grade 12 November average* is partly made up by the other high school mark variables, those variables fluctuate greatly among each other for each student. It is, for example, not uncommon for a student with a high *Mathematics* mark to have a lower *English (2nd)* mark.

Table 8.8 contains the log likelihood p-values for the three new combinations. As expected, they all three have very low log likelihood p-values since they are constructed using the variables that individually have very low log likelihood p-values according to Table 8.5.

Table 8.8: *Log likelihood p-values for each additional combination of Case Study B.*

Combination number	Log likelihood p-value	Combination number	Log likelihood p-value	Combination number	Log likelihood p-value
32	**0.001236**	**33**	$\mathbf{1.31 \times 10^{-15}}$	**34**	$\mathbf{3.46 \times 10^{-16}}$

Based on the above considerations, Combinations 1, 3–11, 17, 23, 25 and 27–34 were chosen for analysis purposes due to the low log likelihood p-values they obtained.

8.3.7 Evaluation of the individual independent variables

The individual independent variables of the first twenty one combinations that remain are analysed in the four subsections of this section.

Their p-values, $\hat{\beta}$-coefficients, ORs, $\hat{\beta}$-CIs, and OR CIs

The p-values, $\hat{\beta}$-coefficients, $\hat{\beta}$-CIs, and OR CIs for each of the independent variables in each of the first twenty one combinations are shown in Tables 8.9–8.29.

For Combination 1 (see Table 8.9), containing the variable *Gender*, it is evident that the category *Female* is significantly different from the category *Male* (the reference class). The positive $\hat{\beta}$-coefficient and OR value of 1.77 indicates that the odds of success of female students are greater than that for males. The widths of the two CIs are also reasonably small, which bodes well for the precision of this result.

Income per family member is the only variable present in Combination 3 (see Table 8.10). By investigating its p-value one can see that it is statistically significant and its small CI width is also a positive sign. Despite this, this variable might not be valuable for predicting tertiary success since its $\hat{\beta}$-coefficient is practically zero and as a result its OR is very close to one.

Combination 4 (see Table 8.11) contains the variable *Study region* with *Gauteng* as its reference class. One can see that all provinces, except *Eastern Cape*, have small p-values. All provinces furthermore have positive $\hat{\beta}$-coefficients, as expected, since the reference class was chosen as the category containing the lowest proportion of students who were successful at tertiary academics.

Table 8.9: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 1 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	402291	0.0702	−0.0941	0.2345	0.3286	1.0727	0.9102	1.2613	0.3541
Female	0.000004	0.5725	0.3296	0.8155	0.4859	1.7727	1.3904	2.2602	0.8698

Table 8.10: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 3 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	3.4×10^{-10}	0.5165	0.3553	0.6778	0.3225	1.6762	1.4266	1.9696	0.5430
Inc/mem	0.010243	−0.0002	−0.0004	−0.0001	0.0004	0.9998	0.9996	0.9999	0.0004

TABLE 8.11: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 4 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.275518	−0.1192	−0.3334	0.0950	0.4285	0.8876	0.7165	1.0997	0.3832
EC	0.239281	0.2463	−0.1639	0.6566	0.8206	1.2793	0.8488	1.9283	1.0795
FS	0.064413	0.4069	−0.0243	0.8381	0.8624	1.5021	0.9759	2.3119	1.3360
KZN	2.1×10^{-11}	1.2025	0.8507	1.5544	0.7037	3.3285	2.3412	4.7322	2.3909
LP	0.067184	0.9947	−0.0704	2.0597	2.1301	2.7038	0.9320	7.8436	6.9115
NW	0.016626	1.0355	0.1880	1.8829	1.6949	2.8165	1.2069	6.5726	5.3658
WC	0.030184	0.3976	0.0381	0.7570	0.7189	1.4882	1.0388	2.1320	1.0931

The category with the lowest p-value, *KwaZulu-Natal*, indicating that it is statistically the most significantly different from *Gauteng*, also has the largest positive $\hat{\beta}$-coefficient, which translates into the largest OR. It may thus be stated that the odds of a student being successful at tertiary level who attended a tertiary institution in *KwaZulu-Natal* over the odds of a student being successful at tertiary level who attended a tertiary institution in *Gauteng* is 3.33. The categories of *Free State*, *KwaZulu-Natal*, and *Western Cape* have the smallest CIs in conjunction with small p-values, which indicates good precision combined with the statistical significance.

Overall it would seem that *Study region* is a good predictor of tertiary success, but further analysis is required to confirm this.

Combination 5 (see Table 8.12) contains the single variable of *Tertiary institution*. The *University of Pretoria* was selected as the reference class, but it could also have been *Stellenbosch University* as they were tied for the category with the smallest proportion of students who were successful. This would explain the extremely high p-value of 99.6% for *Stellenbosch University*.

TABLE 8.12: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 5 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.003598	−0.8232	−1.3774	−0.2690	1.1083	0.4390	0.2522	0.7641	0.5119
CPUT	1.9×10^{-8}	2.5772	1.6784	3.4760	1.7976	13.1605	5.3572	32.3302	26.9731
CUT	0.008445	1.1417	0.2921	1.9912	1.6992	3.1319	1.3392	7.3247	5.9855
DUT	1.9×10^{-9}	2.2855	1.5397	3.0312	1.4915	9.8304	4.6634	20.7226	16.0593
NMMU	0.002247	1.1269	0.4040	1.8497	1.4457	3.0860	1.4979	6.3580	4.8601
NWU	0.001566	1.5163	0.5765	2.4562	1.8797	4.5556	1.7798	11.6604	9.8807
Rhodes	0.779762	0.1301	−0.7815	1.0416	1.8231	1.1389	0.4577	2.8338	2.3761
SMHSU	0.013936	1.8040	0.3661	3.2420	2.8760	6.0741	1.4420	25.5850	24.1429
SU	0.996143	0.0022	−0.8978	0.9022	1.8000	1.0022	0.4075	2.4651	2.0576
TUT	0.000002	2.0889	1.2201	2.9577	1.7376	8.0758	3.3874	19.2532	15.8658
UCT	0.064039	0.8232	−0.0480	1.6944	1.7425	2.2778	0.9531	5.4436	4.4904
UFS	0.030245	0.8624	0.0823	1.6425	1.5601	2.3689	1.0858	5.1680	4.0822
UJ	0.047858	0.6690	0.0063	1.3318	1.3255	1.9524	1.0063	3.7878	2.7815
UWC	0.009459	0.9774	0.2392	1.7155	1.4764	2.6574	1.2702	5.5596	4.2894
UKZN	3.0×10^{-7}	1.7004	1.0502	2.3507	1.3005	5.4764	2.8582	10.4929	7.6347
Wits	0.500890	0.2387	−0.4563	0.9337	1.3901	1.2696	0.6336	2.5439	1.9104

Those tertiary institutions identified as being statistically the most significant according to the p-values are *Cape Peninsula University of Technology*, *Durban University of Technology*, and *University of KwaZulu-Natal*. All three similarly have a medium size CI width, but the OR for *Cape Peninsula University of Technology* and *Durban University of Technology* are approximately twice that of *University of KwaZulu-Natal*.

Besides *Stellenbosch University*, the only other categories that do not show statistical significance are *Rhodes University* and the *University of the Witwatersrand*.

Notice how, although the *University of Cape Town* and the *University of the Witwatersrand* have relatively small p-values, their $\hat{\beta}$-CIs contain zero, indicating that they are not significantly different from the reference class of *University of Pretoria*. The reason for this is that a 95% CI was used, whereas if it were, say, a 90% CI, they would both show up as being significant. This again illustrates why one should not make conclusions about variables and their categories based on a strict α cut-off value.

All of the categories have very similar $\hat{\beta}$-CI widths, which are of a medium size, except for *Sefako Makgatho Health Sciences University*. The reason for this lack of precision may be that there exists a much smaller number of observations within the category of *Sefako Makgatho Health Sciences University* than for any of the other tertiary institutions.

Contained within Combination 6 (see Table 8.13) is the variable *Study field*. As for the previous combination, the $\hat{\beta}$-coefficients are unsurprisingly all positive since the reference class (*Engineering*) was selected as the category containing the smallest proportion of students who were successful at their tertiary studies. According to the p-values, all study fields seem to be significantly different from *Engineering*, with the exception of *Law*. *Commerce* and *Education* furthermore have very small p-values and *Humanities* and *Management* extremely small p-values.

TABLE 8.13: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 6 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.002702	−0.4212	−0.6964	−0.1460	0.5504	0.6563	0.4984	0.8642	0.3658
Arts	0.003397	1.1144	0.3688	1.8600	1.4912	3.0476	1.4459	6.4235	4.9776
Blt Env	0.054108	0.9067	−0.0160	1.8294	1.8455	2.4762	0.9841	6.2304	5.2463
Bus Man	0.002992	2.2930	0.7791	3.8069	3.0278	9.9048	2.1795	45.0125	42.8330
Comm	0.000073	0.7871	0.3981	1.1762	0.7782	2.1971	1.4889	3.2422	1.7532
Edu	0.000102	1.5198	0.7531	2.2866	1.5335	4.5714	2.1235	9.8414	7.7179
Hum	7.0×10^{-8}	1.4650	0.9326	1.9975	1.0649	4.3276	2.5410	7.3704	4.8294
Law	0.374106	0.2728	−0.3288	0.8743	1.2031	1.3136	0.7198	2.3973	1.6775
Man	3.2×10^{-11}	1.8914	1.3328	2.4499	1.1171	6.6286	3.7918	11.5876	7.7957
Med	0.000016	1.1885	0.6480	1.7289	1.0809	3.2821	1.9117	5.6346	3.7228
Sci	0.036464	0.4212	0.0265	0.8159	0.7894	1.5238	1.0269	2.2612	1.2343
Tech	0.076382	0.6444	−0.0683	1.3570	1.4254	1.9048	0.9340	3.8847	2.9507

It would also seem that the category with only the sixth smallest p-value has the largest $\hat{\beta}$-coefficient and thus the largest OR, but unfortunately also the largest CI width, indicating that a large tolerance is associated with the OR. The categories with significant p-values and $\hat{\beta}$-CIs of approximately one or less are *Commerce*, *Humanities*, *Management*, *Medical*, and *Science*.

The last combination containing a single qualitative independent variable is Combination 7 (see Table 8.14) which contains the variable *Qualification type*. The results for *Qualification type* in this case study are similar to those of Case Study A. Again the category of *National diploma* is found to be significantly different to that of *Extended degree* (the reference class). The OR is a healthy 3.5, but note that the CI width is of medium size. Thus, the variable *Qualification type* once again shows potential as a predictor of tertiary success. No pronouncements can be made about the statistical significance of the difference between *Degree* and *Qualification type*.

Table 8.14: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 7 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.547049	−0.1823	−0.7757	0.4111	1.1868	0.8333	0.4604	1.5085	1.0481
Degree	0.527391	0.1976	−0.4153	0.8106	1.2258	1.2185	0.6602	2.2491	1.5890
Nat dip	0.000125	1.2436	0.6081	1.8791	1.2710	3.4680	1.8369	6.5474	4.7104

Tables 8.15–8.17 relate to Combinations 8–10, which contain the three high school average variables. The two variables that obtain statistical significance statuses, according to their p-values, are *Grade 11 November average* with a p-value of approximately 1% and *Grade 12 November average* with a p-value of approximately 0.3%. The variable *Grade 12 June average* performs disappointingly in terms of tertiary success prediction with a p-value larger than 10%. This is not surprising since it achieved a similar log likelihood p-value in Table 8.5.

Interestingly, all three high school average variables have negative $\hat{\beta}$-coefficients, suggesting that an increase in one of these variables leads to a higher tertiary withdrawal rate. This, of course, is counter-intuitive, but agrees with the finding of Case Study A, namely that as the high school marks of students increase, their rates of success at tertiary level decrease.

All three high school averages have small CIs which indicate precision in respect of the $\hat{\beta}$-coefficient estimates. Unfortunately, however, despite all of the interesting findings involving these variables, and especially *Grade 11 November average* and *Grade 12 November average*, all three variables have $\hat{\beta}$-coefficients that are very close to zero. Thus, although one may be confident that a decrease in one of the variables will lead to a decrease in tertiary success, the change in the rate of such success will only be minor.

Table 8.15: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 8 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.009000	2.3089	0.5764	4.0414	3.4650	10.0632	1.7796	56.9037	55.1241
Gr. 11 Nov avg	0.011218	−0.0347	−0.0615	−0.0079	0.0536	0.9659	0.9403	0.9922	0.0518

Table 8.16: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 9 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.099190	1.6271	−0.3071	3.5613	3.8684	5.0891	0.7356	35.2082	34.4726
Gr. 12 June avg	0.121735	−0.0229	−0.0520	0.0061	0.0581	0.9773	0.9493	1.0061	0.0568

TABLE 8.17: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 10 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000328	2.1187	0.9626	3.2748	2.3122	8.3204	2.6186	26.4378	23.8192
Gr. 12 Nov avg	0.002679	−0.0266	−0.0440	−0.0092	0.0347	0.9738	0.9570	0.9908	0.0338

Tables 8.18–8.21 contain the p-values, $\hat{\beta}$-coefficients, OR, and their CIs for the four high school subjects that showed the most potential in terms of tertiary success prediction according to their log likelihood p-values. The first interesting aspect to notice is that, like the three high school average variables, all four high school subject variables have negative $\hat{\beta}$-coefficients, reinforcing the peculiar trend that increasing high school marks tend to indicate a higher withdrawal rate.

The variables *Grade 12 Accounting* and *Grade 12 English* have p-values of approximately 2% and 3%, respectively, as well as small CIs, which indicate good precision of the $\hat{\beta}$-coefficient estimates. Similarly, *Grade 12 Mathematics* and *Grade 12 Physical science* have small CIs and even smaller p-values.

Despite all of the above, all four high school subject variables (*Grade 12 Accounting*, *Grade 12 English*, *Grade 12 Mathematics*, and *Grade 12 Physical science*) unfortunately have $\hat{\beta}$-coefficients that are close to zero and thus ORs close to one. This implies that no matter how precise the $\hat{\beta}$-coefficients and OR estimates are, they only indicate a small slope. This does, however, not take away from the important finding of the negative relationship between the subject marks and tertiary success.

TABLE 8.18: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 11 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000471	2.1582	0.9485	3.3678	2.4193	8.6552	2.5818	29.0150	26.4332
Gr. 12 Accounting	0.021387	−0.0209	−0.0387	−0.0031	0.0356	0.9793	0.9620	0.9969	0.0349

TABLE 8.19: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 17 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.003571	1.4057	0.4601	2.3512	1.8910	4.0782	1.5843	10.4979	8.9136
Gr. 12 English (2nd)	0.028734	−0.0160	−0.0304	−0.0017	0.0287	0.9841	0.9701	0.9983	0.0282

TABLE 8.20: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 23 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000000	1.8474	1.1627	2.5321	1.3694	6.3430	3.1985	12.5793	9.3808
Gr. 12 Mathematics	0.000002	−0.0243	−0.0343	−0.0144	0.0199	0.9760	0.9663	0.9857	0.0195

Table 8.21: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 25 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000208	1.6531	0.7795	2.5267	1.7472	5.2231	2.1804	12.5121	10.3317
Gr. 12 Physical sci	0.000316	−0.0251	−0.0388	−0.0114	0.0273	0.9752	0.9620	0.9886	0.0266

Combination 27 (see Table 8.22) contains the first collection of multiple variables. The variables contained within this combination are *Gender*, *Race*, *Study region*, *Study field*, *Qualification type*, and *Income per family member*. Although there are variations in the values of some of the categories of the five qualitative variables, it is still evident that all of them, except *Race*, have at least one statistically significant category. Keep in mind that the interpretation of the values for each row differs now that there are multiple variables. Consider, as an example, the category *National diploma* of the variable *Qualification type*. In Table 8.14, the interpretation of the OR for this variable was that the odds of a student who pursued a *National diploma* being successful at tertiary level over the odds of student who pursued an *Extended degree* being successful at tertiary level is 3.5. In the multiple logistic regression model, however, the interpretation for the same variable is that the odds of a student who pursued a *National diploma* being successful at tertiary level over the odds of student who pursued an *Extended degree* being successful at tertiary level is 4.6, when keeping the other independent (control) variables constant (that is, keeping them at the same levels and thus neutralising their effects). The minor differences in the OR values of the categories between Tables 8.14 and 8.22 may, in part, be attributed to the change in interpretation.

Table 8.22: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 27 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.020842	−1.2847	−2.3744	−0.1951	2.1793	0.2767	0.0931	0.8228	0.7297
Female	0.020314	0.3608	0.0560	0.6655	0.6095	1.4345	1.0577	1.945	0.8878
African	0.718506	0.1258	−0.5583	0.8099	1.3682	1.1341	0.5722	2.2476	1.6754
EC	0.595617	−0.1389	−0.6521	0.3742	1.0263	0.8703	0.5210	1.4538	0.9329
FS	0.551612	0.1628	−0.3733	0.6989	1.0722	1.1768	0.6885	2.0116	1.3231
KZN	0.000191	0.8176	0.3881	1.2471	0.8590	2.2651	1.4742	3.4802	2.0061
LP	0.093965	0.9487	−0.1615	2.0590	2.2205	2.5825	0.8509	7.8382	6.9873
NW	0.170767	0.7005	−0.3019	1.7029	2.0047	2.0148	0.7394	5.4898	4.7503
WC	0.081890	0.4398	−0.0557	0.9353	0.9909	1.5524	0.9459	2.5480	1.6021
Arts	0.005552	1.3730	0.4026	2.3434	1.9408	3.9472	1.4957	10.4169	8.9211
Blt Env	0.344368	0.5726	−0.6143	1.7595	2.3738	1.7729	0.5410	5.8098	5.2688
Bus Man	0.056951	2.0490	−0.0606	4.1587	4.2192	7.7604	0.9412	63.9853	63.0440
Comm	0.000055	1.0663	0.5481	1.5845	1.0363	2.9046	1.7300	4.8766	3.1467
Edu	0.000023	2.0234	1.0864	2.9605	1.8741	7.5643	2.9636	19.3073	16.3437
Hum	0.000001	1.7945	1.0840	2.5051	1.4211	6.0166	2.9564	12.2445	9.2881
Law	0.023396	0.8646	0.1171	1.6122	1.4951	2.3741	1.1242	5.0136	3.8894
Man	0.000060	1.3861	0.7093	2.0628	1.3535	3.9991	2.0326	7.8680	5.8354
Med	0.000031	1.5298	0.8106	2.2489	1.4383	4.6171	2.2493	9.4777	7.2285
Sci	0.003395	0.7773	0.2572	1.2973	1.0400	2.1755	1.2934	3.6593	2.3659
Tech	0.174270	0.5845	−0.2587	1.4276	1.6863	1.7940	0.7721	4.1687	3.3967
Degree	0.805859	0.0878	−0.6125	0.7881	1.4005	1.0918	0.5420	2.1992	1.6571
Nat dip	0.000076	1.5185	0.7661	2.2708	1.5047	4.5653	2.1515	9.6875	7.5360
Inc/mem	0.308521	−0.0001	−0.0003	0.0001	0.0004	0.9999	0.9997	1.0001	0.0004

The difference in values of *Income per family member* for this combination, as opposed to Combination 3, is that the p-value no longer indicates statistical significance. This is, however, almost irrelevant as once again the $\hat{\beta}$-coefficients are practically zero.

Combination 28 (see Table 8.23) contains all of the variables of Combination 27 with the exception that *Tertiary institution* replaces *Study region*.

TABLE 8.23: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 28 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.021196	−1.5042	−2.7836	−0.2249	2.5587	0.2222	0.0618	0.7986	0.7368
Female	0.026685	0.3630	0.0420	0.6841	0.6421	1.4377	1.0429	1.9820	0.9391
African	0.765607	−0.1148	−0.8693	0.6398	1.5091	0.8916	0.4192	1.8960	1.4768
CPUT	0.003225	1.8942	0.6337	3.1547	2.5209	6.6472	1.8846	23.4448	21.5602
CUT	0.528784	0.4090	−0.8637	1.6816	2.5453	1.5053	0.4216	5.3744	4.9528
DUT	0.060539	1.0496	−0.0465	2.1457	2.1922	2.8566	0.9546	8.5484	7.5938
NMMU	0.212637	0.6066	−0.3473	1.5604	1.9077	1.8341	0.7066	4.7609	4.0543
NWU	0.072923	1.0109	−0.0939	2.1158	2.2097	2.7481	0.9103	8.2962	7.3858
Rhodes	0.989934	−0.0069	−1.0819	1.0681	2.1500	0.9931	0.3389	2.9098	2.5708
SMHSU	0.057013	1.5436	−0.0460	3.1332	3.1793	4.6814	0.9550	22.9478	21.9928
SU	0.970257	0.0219	−1.1318	1.1757	2.3075	1.0222	0.3225	3.2404	2.9179
TUT	0.095026	1.0253	−0.1784	2.2289	2.4073	2.7878	0.8366	9.2899	8.4533
UCT	0.036122	1.0983	0.0711	2.1256	2.0546	2.9992	1.0737	8.3781	7.3044
UFS	0.099229	0.7733	−0.1460	1.6926	1.8386	2.1668	0.8641	5.4334	4.5692
UJ	0.427961	0.3450	−0.5080	1.1979	1.7059	1.4119	0.6017	3.3132	2.7115
UWC	0.120809	0.7359	−0.1938	1.6657	1.8595	2.0874	0.8238	5.2893	4.4655
UKZN	0.000257	1.4561	0.6752	2.2369	1.5617	4.2890	1.9645	9.3643	7.3999
Wits	0.189495	0.5446	−0.2689	1.3580	1.6269	1.7239	0.7642	3.8886	3.1244
Arts	0.005767	1.4445	0.4190	2.4700	2.0510	4.2397	1.5204	11.8223	10.3019
Blt Env	0.440554	0.4755	−0.7329	1.6840	2.4169	1.6089	0.4805	5.3869	4.9064
Bus Man	0.973216	15.0103	−861.23	891.24	1752.5	∞	0.0000	∞	∞
Comm	0.000085	1.1413	0.5722	1.7104	1.1382	3.1308	1.7722	5.5310	3.7589
Edu	0.000161	1.8728	0.9000	2.8456	1.9456	6.5062	2.4595	17.2113	14.7519
Hum	0.000011	1.6538	0.9170	2.3906	1.4736	5.2267	2.5017	10.9199	8.4182
Law	0.036224	0.8359	0.0537	1.6181	1.5644	2.3068	1.0551	5.0434	3.9883
Man	0.000451	1.2961	0.5720	2.0203	1.4483	3.6551	1.7718	7.5402	5.7685
Med	0.000534	1.3465	0.5844	2.1085	1.5241	3.8438	1.7939	8.2361	6.4422
Sci	0.006793	0.7597	0.2096	1.3098	1.1002	2.1376	1.2332	3.7053	2.4721
Tech	0.199722	0.5591	−0.2954	1.4136	1.7091	1.7491	0.7442	4.1109	3.3667
Degree	0.758679	0.1137	−0.6116	0.8390	1.4506	1.1204	0.5425	2.3141	1.7716
Nat dip	0.007230	1.3691	0.3701	2.3681	1.9980	3.9319	1.4479	10.6774	9.2295
Inc/mem	0.276394	−0.0001	−0.0003	0.0001	0.0004	0.9999	0.9997	1.0001	0.0004

With respect to the p-values of the variables and categories found in both Combination 27 and 28, it would appear that the majority remained at approximately the same level, with two exceptions. The p-value of the *National diploma* category changed from 0.008% to 0.7%. Although for Combinations 27 and 28 this category achieves p-values that are statistically significant, this shift of two decimal points should still be noted. This phenomenon is possibly due to the fact that high, but not severe, multicollinearity was identified between *Qualification type* and *Tertiary institution*. The second exception is *Business management*, for which the p-value has increased from 6% to 97%, and its $\hat{\beta}$-coefficient and OR have increased drastically. This change is likely due to multicollinearity between it and one or more of the categories of *Tertiary institution*.

The other minor changes in the p-values and $\hat{\beta}$-coefficients of the categories may, in part, be attributed to slight multicollinearity between the categories of the different variables, but is also largely due to the change in interpretation of the values as a result of the inclusion of control variables.

Despite some more dramatic differences between the results of Combinations 27 and 28, specifically within the categories, the overall picture remains the same, namely that the five variables *Gender*, *Study region*, *Tertiary institution*, *Study field*, and *Qualification type* would appear to be good predictors of tertiary success, while *Race* and *Income per family member* do not exhibit similar potential.

The variables within Combination 29 (see Table 8.24) are identical to those in Combination 27, with the addition of *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*.

TABLE 8.24: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 29 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.002202	−8.3179	−13.6430	−2.9928	10.6502	0.0002	0.0000	0.0501	0.0501
Female	0.506838	−0.2362	−0.9336	0.4612	1.3948	0.7896	0.1579	3.956	3.8030
African	0.761657	0.2518	−1.3752	1.8787	3.2539	1.2863	0.2528	6.5450	6.2922
EC	0.945638	−0.0418	−1.2419	1.1584	2.4004	0.9591	0.2888	3.1849	2.8961
FS	0.439772	−0.6253	−2.2118	0.9611	3.1729	0.5351	0.1095	2.6146	2.5051
KZN	0.207891	0.5821	−0.3238	1.4881	1.8119	1.7898	0.7234	4.4287	3.7053
LP	0.992091	−14.4259	−2867.0	2838.1	5705.1	0.0000	0.0000	∞	∞
NW	0.278224	1.1260	−0.9093	3.1614	4.0707	3.0834	0.4028	23.6038	23.2010
WC	0.772151	0.1826	−1.0531	1.4182	2.4713	1.2003	0.3489	4.1297	3.7809
Arts	0.003291	3.0035	1.0006	5.0064	4.0057	20.156	2.720	149.361	146.641
Blt Env	0.004698	4.0738	1.2495	6.8982	5.6487	58.7826	3.4886	990.496	987.007
Bus Man	0.987348	16.1036	−1974.2	2006.5	3980.7	∞	0.0000	∞	∞
Comm	0.000793	2.3155	0.9628	3.6683	2.7054	10.1305	2.6191	39.1833	36.564
Edu	0.000001	5.8007	3.5226	8.0788	4.5562	330.536	33.873	3 225.445	3 191.572
Hum	0.000000	5.2328	3.3491	7.1166	3.7675	187.318	28.476	1 232.195	1 203.719
Law	0.012974	2.0323	0.4291	3.6355	3.2065	7.6316	1.5358	37.9224	36.3865
Man	0.022349	2.7421	0.3894	5.0947	4.7053	15.5188	1.4761	163.154	161.677
Med	0.000701	3.0022	1.2659	4.7384	3.4725	20.1291	3.5464	114.253	110.707
Sci	0.000755	2.1397	0.8948	3.3846	2.4898	8.4971	2.4469	29.5070	27.0601
Tech	0.002144	2.7845	1.0066	4.5625	3.5559	16.1921	2.7362	95.8206	93.0844
Degree	0.396250	−0.6379	−2.1118	0.8359	2.9478	0.5284	0.1210	2.3070	2.1859
Nat dip	0.000943	2.8204	1.1488	4.4921	3.3433	16.7844	3.1544	89.3100	86.1556
Inc/mem	0.307848	0.0002	−0.0002	0.0006	0.0008	1.0002	0.9998	1.0006	0.0008
Gr.12 Nov avg	0.000724	0.1211	0.0509	0.1913	0.1405	1.1288	1.0522	1.2109	0.1587
Gr.12 June avg	0.238563	−0.0385	−0.1025	0.0255	0.1281	0.9622	0.9026	1.0259	0.1233
Gr.11 Nov avg	0.572129	−0.0131	−0.0587	0.0324	0.0911	0.9870	0.9430	1.0330	0.0899

Two interesting observations are the following: For the first time *Gender* is not statistically significant any more. This is odd, as nearly no collinearity is detected between the variables *Gender* and *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*. A second peculiar finding is that none of the categories of *Study region* is statistically significant (keep in mind, however, that only *KwaZulu-Natal* obtained a p-value of less than 5% in Combination 27).

The above two findings may simply be due to the interpretation of the p-values, where the additional control variables have to be taken into consideration in the latter case. Also, consider the sample size. Whereas the combination analysis for Combinations 27 made use of 908 observations, that for Combination 29 only considered 254 observations. Another plausible explanation for the second finding above may be that multicollinearity is present between the two variables *Grade 12 November average* and *Study region*. Although such multicollinearity may be above average, it is not severe as this would have been flagged in an earlier step.

Of the three additional high school average variables, only *Grade 12 November average* exhibits statistical significance. Unlike before, however, its OR is slightly larger than one, but still not significantly larger.

As mentioned above, the results for specific combinations should first be verified by the other analysis methods (random forest variable importance and CART) before reporting back due to the above average multicollinearity present and often small sample sizes.

Combination 30 (see Table 8.25) is identical to Combination 28, with the addition of the variable *Grade 12 November average*.

Table 8.25: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 30 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000036	−4.9243	−7.2602	−2.5884	4.6719	0.0073	0.0007	0.0751	0.0744
Female	0.037369	0.3800	0.0222	0.7377	0.7155	1.4623	1.0225	2.0912	1.0687
African	0.805808	0.1049	−0.7311	0.9408	1.6719	1.1105	0.4814	2.5620	2.0807
CPUT	0.002195	2.1996	0.7919	3.6073	2.8154	9.0212	2.2076	36.8650	34.6574
CUT	0.476770	0.5133	−0.9007	1.9273	2.8280	1.6708	0.4063	6.8709	6.4646
DUT	0.046442	1.2473	0.0197	2.4749	2.4553	3.4809	1.0199	11.8811	10.8612
NMMU	0.030220	1.2382	0.1184	2.3580	2.2396	3.4494	1.1257	10.5696	9.4439
NWU	0.045292	1.2879	0.0270	2.5488	2.5218	3.6252	1.0274	12.7921	11.7647
Rhodes	0.770254	0.1761	−1.0058	1.3581	2.3639	1.1926	0.3657	3.8887	3.5230
SMHSU	0.985450	17.6032	−1874.3	1909.5	3783.9	∞	0.0000	∞	∞
SU	0.763413	−0.2068	−1.5534	1.1398	2.6932	0.8132	0.2115	3.1261	2.9145
TUT	0.040561	1.4991	0.0644	2.9337	2.8693	4.4775	1.0665	18.7974	17.7309
UCT	0.060570	1.0943	−0.0486	2.2373	2.2859	2.9872	0.9526	9.3677	8.4151
UFS	0.015629	1.2938	0.2448	2.3427	2.0979	3.6465	1.2774	10.4093	9.1318
UJ	0.200928	0.6279	−0.3344	1.5902	1.9246	1.8737	0.7158	4.9046	4.1889
UWC	0.028450	1.1824	0.1247	2.2401	2.1154	3.2622	1.1328	9.3943	8.2615
UKZN	0.000351	1.6321	0.7372	2.5271	1.7899	5.1147	2.0900	12.5166	10.4266
Wits	0.129914	0.7068	−0.2079	1.6214	1.8294	2.0274	0.8123	5.0604	4.2481
Arts	0.000650	2.0801	0.8844	3.2757	2.3912	8.0049	2.4216	26.4607	24.0391
Blt Env	0.184008	0.9417	−0.4476	2.3310	2.7786	2.5643	0.6392	10.2882	9.6490
Bus Man	0.981937	16.4236	−1405.4	1438.2	2843.6	∞	0.0000	∞	∞

. . . continued

. . . continued

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Comm	0.000038	1.3307	0.6970	1.9644	1.2674	3.7836	2.0077	7.1303	5.1226
Edu	0.000002	2.7662	1.6368	3.8956	2.2588	15.8986	5.1389	49.1871	44.0482
Hum	$9.4{\times}10^{-8}$	2.3217	1.4693	3.1742	1.7049	10.1934	4.3462	23.9072	19.5611
Law	0.099345	0.7440	−0.1408	1.6289	1.7697	2.1044	0.8686	5.0982	4.2295
Man	0.000029	1.7839	0.9484	2.6194	1.6710	5.9531	2.5817	13.7275	11.1458
Med	0.000109	1.6832	0.8308	2.5357	1.7049	5.3830	2.2952	12.6252	10.3300
Science	0.000738	1.0525	0.4413	1.6637	1.2224	2.8649	1.5548	5.2791	3.7243
Tech	0.064388	0.8755	−0.0523	1.8032	1.8555	2.4000	0.9490	6.0691	5.1201
Degree	0.262769	−0.4528	−1.2452	0.3397	1.5849	0.6359	0.2879	1.4045	1.1166
Nat dip	0.036941	1.1661	0.0707	2.2615	2.1909	3.2095	1.0732	9.5979	8.5247
Inc/mem	0.400990	−0.0001	−0.0003	0.0001	0.0004	0.9999	0.9997	1.0001	0.0004
Gr.12 Nov avg	0.000738	0.0470	0.0197	0.0743	0.0546	1.0481	1.0199	1.0771	0.0572

Once again *Business management* has an extremely high p-value and a large OR. It is either truly not statistically significant at all, or more likely correlated with one of the other variables or categories, but it is not clear which of these is indeed the case. This has, however, until now not hindered the interpretation of results as the variable *Study field* is considered as a whole. It is likely also a similar situation for the case of *Sefako Makgatho Health Sciences University*.

Two other interesting aspects to note is that *Female* is again significantly different from *Male*, while keeping the remaining control variables constant. It would thus seem that the results of Combination 29, which indicated that *Female* is not significantly different from *Male*, influenced by one of the variables Combination 29, was potentially misleading. Again *Grade 12 November average* produces a small p-value, backed by a narrow CI, but with a $\hat{\beta}$-coefficient very close to zero.

At this stage it would appear, barring the peculiar results influenced by multicollinearity and other factors mentioned above, that generally the most influential variables are *Gender*, *Study region*, *Tertiary institution*, *Study field*, and *Qualification type*, while *Grade 12 November average* may also be, but confirmation of this will require further analysis.

Combination 31 (see Table 8.26) contains only the three high school average marks *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*. Of these three variables, it is interesting to note that only *Grade 11 November average* exhibits a relatively small p-value, along with a narrow $\hat{\beta}$-CI, but again the $\hat{\beta}$-coefficient is in close proximity to zero. Keep in mind, however, that the sample of this combination is relatively small (293 observations) and so this finding will have to be corroborated by those of the other combinations.

Combination 32 (see Table 8.27) consists of the same variables as Combination 31, with the addition of *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*. This time none of the three high school average variables shows any statistical significance, and only *Grade 12 Mathematics* and *Grade 12 Physical science* achieve relatively small p-values.

It is interesting to note that the trend of negative $\hat{\beta}$-coefficients for high school subjects continues. As with the four high school subject variables considered up to now, however, *Grade 12 Mathematics* and *Grade 12 Physical science* appear to have narrow CIs, indicating precision, together with the relatively small p-values, but again their $\hat{\beta}$-coefficients are close to zero. These small $\hat{\beta}$-coefficients, as before, translate into small ORs.

Table 8.26: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 31 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.004795	3.9978	1.2197	6.7759	5.5562	54.4798	3.3862	876.5	873.1
Gr.12 Nov avg	0.393758	−0.0177	−0.0583	0.0230	0.0813	0.9825	0.9434	1.0232	0.0799
Gr.12 June avg	0.680752	−0.0089	−0.0514	0.0335	0.0849	0.9911	0.9499	1.0341	0.0842
Gr.11 Nov avg	0.064504	−0.0336	−0.0693	0.0020	0.0713	0.9669	0.9331	1.0020	0.0690

Table 8.27: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 32 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.033693	4.2460	0.3274	8.1646	7.8372	69.8264	1.3874	3514.4	3513.0
Gr.12 Nov avg	0.269414	0.0740	−0.0573	0.2053	0.2626	1.0768	0.9443	1.2279	0.2836
Gr.12 June avg	0.510124	−0.0192	−0.0761	0.0378	0.1140	0.9810	0.9267	1.0386	0.1119
Gr.11 Nov avg	0.153756	−0.0319	−0.0757	0.0119	0.0876	0.9686	0.9271	1.0120	0.0849
English (2nd)	0.648931	−0.0146	−0.0775	0.0483	0.1258	0.9855	0.9254	1.0495	0.1241
Mathematics	0.070472	−0.0348	−0.0726	0.0029	0.0755	0.9658	0.9300	1.0029	0.0729
Physical sci	0.056541	−0.0432	−0.0876	0.0012	0.0888	0.9577	0.9161	1.0012	0.0851

Combination 33 (see Table 8.28) contains the nine variables *Gender*, *Family income per member*, *Study region*, *Study field*, *Qualification type*, *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*. As was previously seen in Combination 29, the category *Female* is again not statistically significant. It would appear that correlation might be present between this category and one of the high school subject variables.

With the exception of the categories *North West* and *Arts*, which might be influenced by multicollinearity, the overall significance of the remaining qualitative variables *Study region*, *Study field* and *Qualification type* appear to be influential in respect of predicting tertiary success.

All four the high school academic variables (*Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*) perform poorly, with none of them achieving a p-value of less than 24%.

Table 8.28: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 33 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.007129	−3.7866	−6.5447	−1.0284	5.5163	0.0227	0.0014	0.3576	0.3561
Female	0.152633	0.3692	−0.1367	0.8752	1.0119	1.4466	0.8722	2.3993	1.5271
EC	0.864886	−0.0724	−0.9057	0.7610	1.6668	0.9302	0.4042	2.1405	1.7362
FS	0.986702	−0.0076	−0.9061	0.8909	1.7970	0.9924	0.4041	2.4372	2.0331
KZN	0.058096	0.6380	−0.0219	1.2979	1.3198	1.8927	0.9784	3.6617	2.6833
LP	0.168132	1.1177	−0.4718	2.7072	3.1790	3.0579	0.6239	14.9873	14.3634
NW	0.985298	17.2476	−1817.2	1851.7	3667.0	∞	0.0000	∞	∞
WC	0.086387	0.5566	−0.0796	1.1927	1.2723	1.7447	0.9235	3.2961	2.3726
Arts	0.986998	18.0928	−2157.9	2194.1	4352.0	∞	0.0000	∞	∞
Blt Env	0.095576	1.4179	−0.2495	3.0854	3.3349	4.1286	0.7792	21.8754	21.0962
Comm	0.000364	1.8361	0.8267	2.8456	2.0189	6.2723	2.2857	17.2120	14.9262

. . . continued

. . . continued

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Edu	0.000730	3.2100	1.3476	5.0725	3.7249	24.7800	3.8481	159.5708	155.7227
Hum	0.000058	2.7818	1.4254	4.1382	2.7128	16.1483	4.1595	62.6924	58.5330
Law	0.032411	2.2600	0.1895	4.3306	4.1411	9.5834	1.2086	75.9899	74.7813
Man	0.129064	1.8779	−0.5470	4.3028	4.8498	6.5395	0.5787	73.9033	73.3247
Med	0.000005	2.2324	1.2778	3.1869	1.9091	9.3219	3.5889	24.2130	20.6242
Sci	0.000563	1.2649	0.5461	1.9837	1.4376	3.5427	1.7265	7.2695	5.5430
Tech	0.014410	1.3973	0.2781	2.5166	2.2385	4.0443	1.3206	12.3860	11.0655
Degree	0.952123	−0.0369	−1.2406	1.1669	2.4075	0.9638	0.2892	3.2119	2.9227
Nat dip	0.000138	2.4526	1.1914	3.7138	2.5224	11.6189	3.2918	41.0110	37.7192
Inc/mem	0.738684	0.0000	−0.0003	0.0002	0.0005	1.0000	0.9997	1.0002	0.0005
Gr.12 Nov avg	0.309081	0.0377	−0.0350	0.1104	0.1453	1.0384	0.9657	1.1167	0.1510
English (2nd)	0.433729	−0.0162	−0.0567	0.0243	0.0810	0.9839	0.9449	1.0246	0.0798
Math	0.241113	0.0146	−0.0098	0.0389	0.0487	1.0147	0.9903	1.0397	0.0494
Phy sci	0.626240	−0.0077	−0.0385	0.0232	0.0617	0.9924	0.9622	1.0235	0.0613

The final combination, Combination 34 (see Table 8.29), contains the variables *Gender*, *Income per family member*, *Tertiary institution*, *Study field*, *Qualification type*, *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*. This combination is identical to Combination 33, except that *Tertiary institution* replaces *Study region*.

Again the category *Female* does not have a small p-value, which indicates that it is not significantly correlated with one of the high school subjects. Also, as in a previous combination, *Sefako Makgatho Health Sciences University* and *Arts* appear to be influenced by multicollinearity.

With the exception of *Grade 12 November average*, which now exhibits a relatively small statistical significance, but still with a small $\hat{\beta}$-coefficient, all of the values associated with the remaining variables and categories appear to be very similar to those of Combination 33.

TABLE 8.29: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 34 of Case Study B.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000710	−5.7234	−9.0366	−2.4101	6.6265	0.0033	0.0001	0.0898	0.0897
Female	0.151201	0.4044	−0.1478	0.9566	1.1044	1.4984	0.8626	2.6028	1.7402
CPUT	0.004200	3.6412	1.1483	6.1341	4.9858	38.1374	3.1528	461.3278	458.1751
CUT	0.613557	0.5039	−1.4518	2.4595	3.9113	1.6551	0.2342	11.6993	11.4652
DUT	0.226407	1.0575	−0.6559	2.7710	3.4269	2.8793	0.5190	15.9746	15.4556
NMMU	0.179259	1.1221	−0.5154	2.7596	3.2750	3.0712	0.5972	15.7930	15.1958
NWU	0.004114	3.5710	1.1317	6.0102	4.8785	35.5512	3.1010	407.5729	404.4719
Rhodes	0.305695	0.8889	−0.8120	2.5899	3.4019	2.4325	0.4440	13.3280	12.8841
SMHSU	0.985090	18.0419	−1874.1	1910.2	3784.4	∞	0.0000	∞	∞
SU	0.429461	0.6633	−0.9821	2.3086	3.2907	1.9411	0.3745	10.0605	9.6860
TUT	0.057886	1.9380	−0.0648	3.9407	4.0055	6.9446	0.9373	51.4549	50.5176
UCT	0.157243	1.0900	−0.4204	2.6004	3.0209	2.9743	0.6568	13.4695	12.8128

. . . continued

. . . continued

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
UFS	0.142405	1.2169	−0.4090	2.8428	3.2519	3.3767	0.6643	17.1643	16.5001
UJ	0.139450	1.0374	−0.3384	2.4132	2.7516	2.8218	0.7129	11.1698	10.4569
UWC	0.004498	2.1195	0.6573	3.5818	2.9245	8.3271	1.9295	35.9377	34.0082
UKZN	0.001088	2.1513	0.8605	3.4420	2.5815	8.5957	2.3644	31.2498	28.8854
Wits	0.070028	1.1563	−0.0946	2.4072	2.5018	3.1782	0.9097	11.1031	10.1934
Arts	0.986879	18.1737	−2147.7	2184.0	4331.7	∞	0.0000	∞	∞
Blt Env	0.052334	1.6849	−0.0170	3.3869	3.4039	5.3920	0.9831	29.5728	28.5897
Comm	0.000655	1.8523	0.7870	2.9176	2.1307	6.3745	2.1967	18.4974	16.3007
Edu	0.000448	3.4486	1.5230	5.3741	3.8512	31.4551	4.5858	215.7556	211.1698
Hum	0.000152	2.7152	1.3101	4.1203	2.8102	15.1072	3.7065	61.5750	57.8685
Law	0.057761	2.1143	−0.0696	4.2981	4.3677	8.2834	0.9328	73.5598	72.6270
Man	0.064866	2.5280	−0.1558	5.2118	5.3676	12.5282	0.8557	183.4180	182.5623
Med	0.000188	1.9690	0.9358	3.0022	2.0664	7.1633	2.5492	20.1295	17.5803
Sci	0.000590	1.3512	0.5805	2.1220	1.5415	3.8622	1.7869	8.3476	6.5607
Tech	0.012428	1.4568	0.3146	2.5991	2.2845	4.2923	1.3697	13.4512	12.0815
Degree	0.946923	−0.0425	−1.2938	1.2088	2.5026	0.9584	0.2742	3.3494	3.0752
Nat dip	0.005163	2.3301	0.6971	3.9630	3.2659	10.2788	2.0080	52.6166	50.6086
Inc/mem	0.735912	0.0000	−0.0003	0.0002	0.0006	1.0000	0.9997	1.0002	0.0006
Gr.12 Nov avg	0.084088	0.0688	−0.0093	0.1468	0.1560	1.0712	0.9908	1.1581	0.1673
English (2nd)	0.172712	−0.0302	−0.0736	0.0132	0.0868	0.9702	0.9290	1.0133	0.0843
Math	0.337899	0.0132	−0.0138	0.0401	0.0539	1.0133	0.9863	1.0409	0.0546
Phy sci	0.497767	−0.0112	−0.0435	0.0211	0.0646	0.9889	0.9575	1.0213	0.0639

In summary, based on all of the above results, it may cautiously and preliminarily be concluded that it appears as if the variables *Study region*, *Tertiary institution*, *Qualification type*, and *Study field* may be used to predict tertiary success of students. The variables *Gender*, *Grade 12 November average*, *Grade 12 Mathematics*, and *Grade 12 Physical science* might also be used in this respect, but further investigation is required for these variables.

Random forest variable importance scores should be able to shed further light on the findings of the case study produced thus far. The signs of the relationships, *i.e.* whether or not presence of a certain variable will result in an increased or decreased probability of success at tertiary level, will also be considered again when CART plots and bar plots are generated for the various logistic regression models.

Random forest variable importance scores

The second step in evaluating the individual variables and their categories was to draw up random forest variable importance plots.

Combinations that are applicable in this context are those that contain more than one variable or a single qualitative variable, that is Combinations 1, 4, 5, 6, 7, and 27–34. Variable importance plots for these combinations may be found in Figures 8.7–8.17. When employing random forests it is possible to include all the categories of variables (reference categories are not applicable).

The random forest variable importance score of Combination 1, containing the two categories *Female* and *Male* of the variable *Gender* is shown in Figure 8.7. The two categories display near identical importance scores, but represent disappointingly low importance. It is still difficult to claim conclusively that the variable *Race* has no predictive power in terms of tertiary academic success as its potential in the logistic regression fluctuated.

Combination 4 (see Figure 8.8), contains the categories of the variable *Study region*. The two categories that stand out are *Gauteng* and *KwaZulu-Natal*. The importance of *Gauteng* is expected since it was previously chosen as the reference class on the basis of being the category containing the observations with the lowest tertiary success rate. The importance of *KwaZulu-Natal* agrees with that of the previous findings in which that province achieved the smallest p-value, the $\hat{\beta}$-coefficient most different from zero, and the narrowest CIs in comparison to the other categories.

The random forest variable importance scores of the sixteen categories of the variable *Tertiary institution*, the only variable included in Combination 5, may be viewed in Figure 8.9. The categories *Cape Peninsula University of Technology*, *Durban University of Technology*, *Tshwane University of Technology*, *University of Pretoria* (the reference class for logistic regression), *University of KwaZulu-Natal*, and *University of the Witwatersrand* all achieve above average importance scores. Overall, the variable *Tertiary institution* again appears to be a valuable variable with respect to predicting tertiary success.

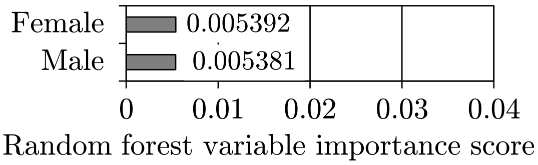


Figure 8.7: *Random forest variable importance scores for Combination 1 of Case Study B.*

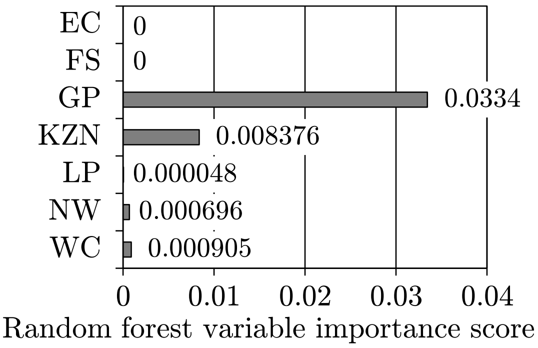


Figure 8.8: *Random forest variable importance scores for Combination 4 of Case Study B.*

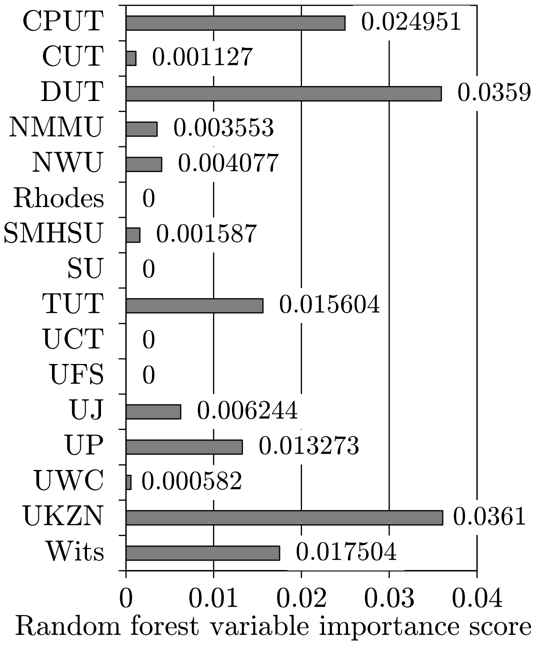


Figure 8.9: *Random forest variable importance scores for Combination 5 of Case Study B.*

Combination 6 (see Figure 8.10) contains the twelve categories of the variable *Study field*. The category shown to be the most important in terms of its variable importance scores is *Engineering*, which was also the reference class for logistic regression. Although this one category exhibits a high importance score, it requires further investigation in order to conclude whether or not the entire variable will be a good predictor of tertiary success. Based on the logistic regression models, however, it is anticipated that it will be shown as such.

The importance values of the next combination are shown in Figure 8.11, and contains the three categories of the variable *Qualification type*. The category *National diploma* is shown to achieve the largest importance value. For this variable, however, the category that was used as the reference class for logistic regression, namely *Extended degree*, does not achieve a high importance value. This is likely due to the small number of observations that are present in this category. Overall, however, the variable *Qualification type* again appears to be influential with respect to predicting tertiary success.

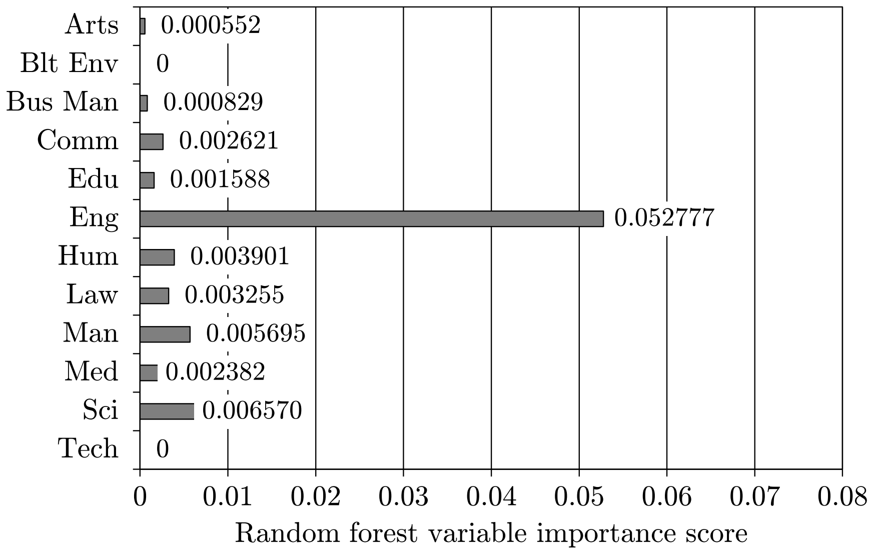


FIGURE 8.10: *Random forest variable importance scores for Combination 6 of Case Study B.*

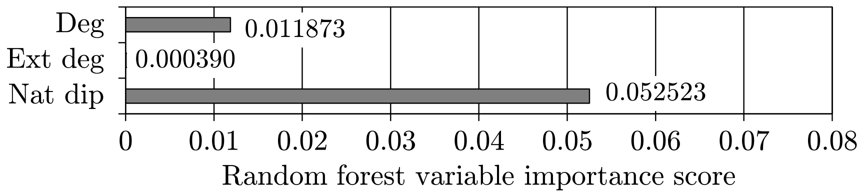


FIGURE 8.11: *Random forest variable importance scores for Combination 7 of Case Study B.*

Figures 8.12–8.13 contain the combined variables of Combinations 27 and 28. For the model of Figure 8.12, each category of each variable is transformed to a dummy column or variable, as has been done with the random forest variable importance score plots up to this point.

For the model of Figure 8.13, the values of each variable were contained within a single column, and so only a single importance score is shown for each variable in this case. This is informative as it shows the importance score of the qualitative variables as a whole (as opposed to only those of their categories).

From the variable importance scores of Combinations 27 and 28, combined with dummy variables and shown in Figure 8.12, it is clear that variables which achieve high importance scores are *Study region*, *Tertiary institution*, *Study field*, and *Qualification type*. The specific categories of these qualitative variables that achieved high scores in the previous combination are again those with the highest scores in this collection of Combinations 27 and 28. Disappointing importance scores are associated with the categories *Female* and *Male* of *Gender*, all of *Race*, and the variable *Income per family member*.

From the variable importance scores of Combinations 27 and 28, which do not utilise dummy variables and which are shown in Figure 8.13, it would appear that the importance scores correspond to those of Figure 8.12, with the four most important variables being *Study region*, *Tertiary institution*, *Study field*, and *Qualification type*.

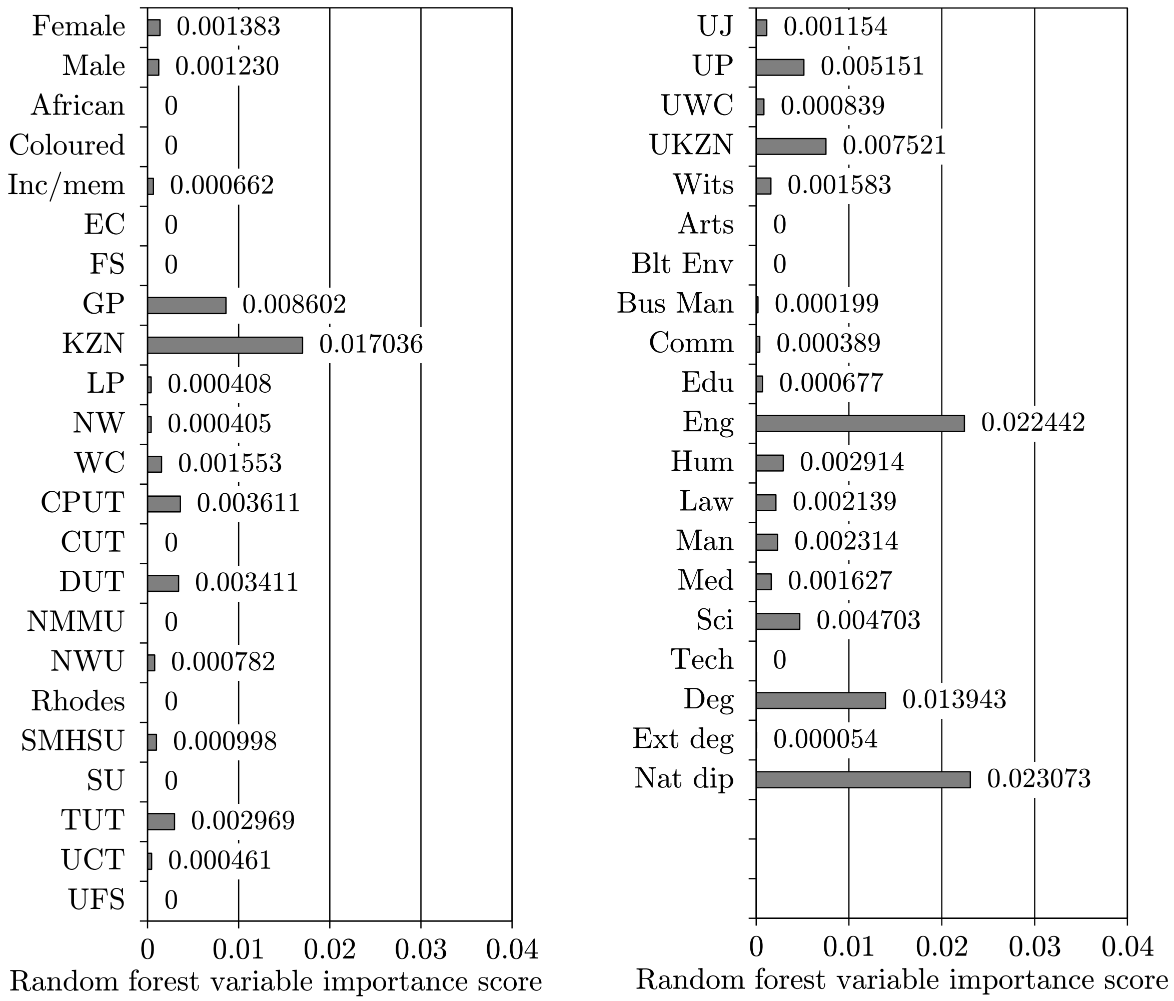


FIGURE 8.12: *Random forest variable importance scores for Combinations 27 and 28 together, including dummy variables, of Case Study B.*

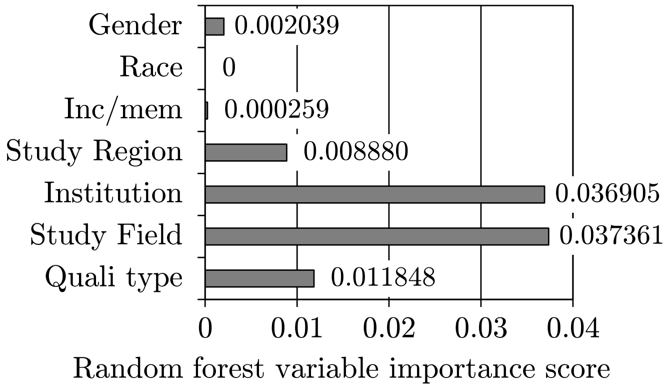


FIGURE 8.13: *Random forest variable importance scores for Combinations 27 and 28 together of Case Study B.*

Figures 8.14–8.15 both contain the variable importance scores for the variables of Combinations 29 and 30 combined, but Figure 8.14 also contains the variable importance scores for each category of qualitative variable, while the variable importance scores shown in Figure 8.15 are only applicable to the variables as a whole.

The scores shown in Figure 8.14 are very similar to those of Figure 8.12, with the exception of *Humanities* which now achieves a higher score. This change may well be a result of the change in sample size, as Combinations 29 and 30 together have approximately only a third of the observations available to Combinations 27 and 28 combined. Of the three new variables included in Combinations 29 and 30, namely *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*, only *Grade 12 June average* produces an importance score slightly above zero, but this score is very small relative to the other scores.

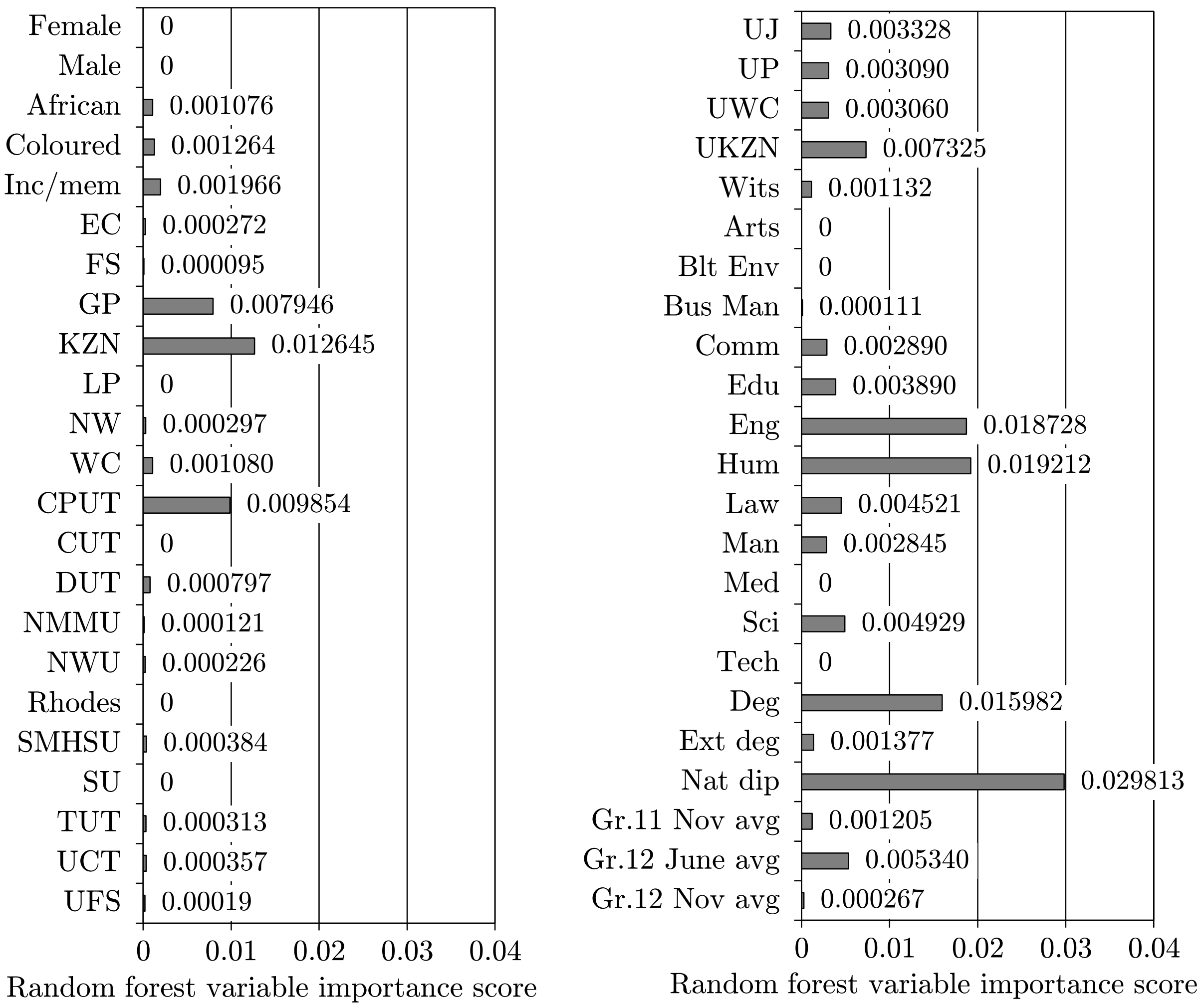


Figure 8.14: *Random forest variable importance scores for Combinations 29 and 30 together, including dummy variables, of Case Study B.*

Figure 8.15, containing the variable importance scores of Combinations 29 and 30 combined without dummy variables, exhibits remarkable similarities to Figure 8.13, containing the scores of Combinations 27 and 28 combined. The difference is that the variables *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average* are also included in Combinations 29 and 30 together, but as shown in Figure 8.14, all of these variables, except *Grade 12 June average*, produce importance scores that are only slightly larger than zero.

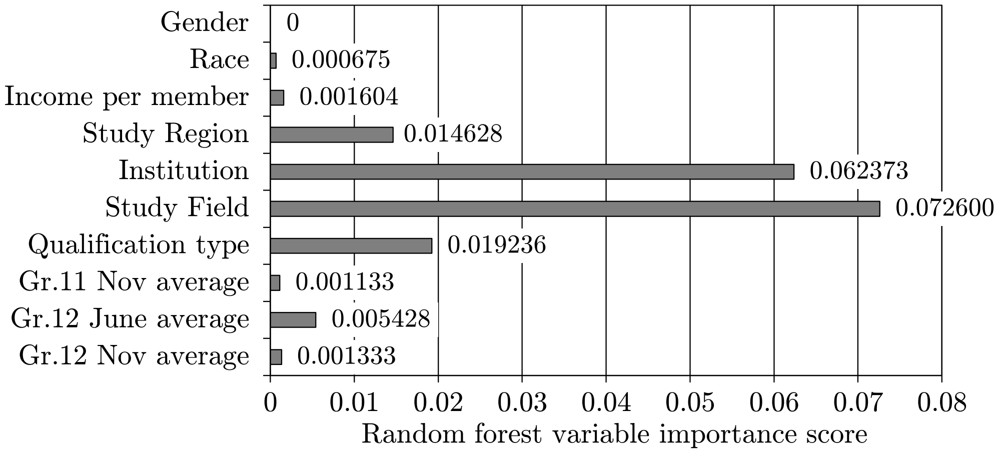


FIGURE 8.15: *Random forest variable importance scores for Combinations 29 and 30 together of Case Study B.*

The next combination should have been Combination 31, containing the three high school average variables, but for that combination none of the three variables produced a variable importance score larger than zero. This does not bode well for this combination and its variables, but it should be taken into consideration that the sample size was only 294 observations.

Combination 32 (see Figure 8.16) contains the six quantitative variables *Grade 11 November average*, *Grade 12 June average*, *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*.

It would seem that *Grade 12 June average* achieves the largest importance score, slightly more than the score associated with it in the previous random forest variable importance plots. This, however, stands in contradiction to the logistic regression model of Combination 32 which only identified *Grade 12 Mathematics* and *Grade 12 Physical science* as slightly significant, as may be observed in this importance score plot as well.

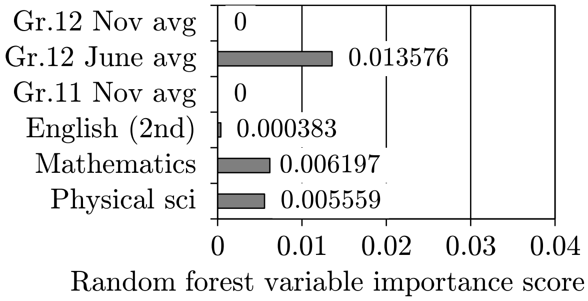


FIGURE 8.16: *Random forest variable importance scores for Combination 32 of Case Study B.*

The last two random forest variable importance plots relate to a model combining the variables of Combinations 33 and 34, that is *Gender*, *Income per family member*, *Study region Tertiary institution*, *Study field*, *Qualification type*, *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics* and *Grade 12 Physical science*.

The first, shown in Figure 8.17, contains importance values for all of the above variables without using dummy columns for the qualitative variables. The scores for variables previously considered are similar to those seen in previous combinations (such as Figures 8.13 and 8.15), with *Study region Tertiary institution*, *Study field*, and *Qualification type* achieving high scores, and *Gender* and *Income per family member* not. Of the three high school variables, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*, none obtains a high score.

Figure 8.18 also contains the combined variables of Combinations 33 and 34, but using dummy columns to produce importance scores for each category of the qualitative variables. As in Figure 8.17, no significant differences are observed between the scores in Figures 8.12, 8.14 and 8.18. Once again no high scores are obtained for the three high school variables *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*.

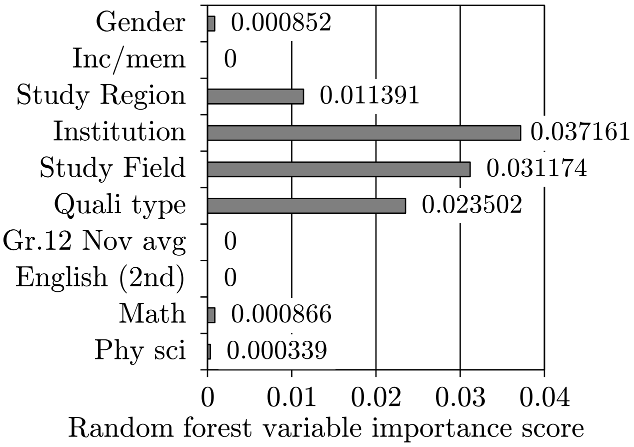


FIGURE 8.17: *Random forest variable importance scores for Combinations 33 and 34 together of Case Study B.*

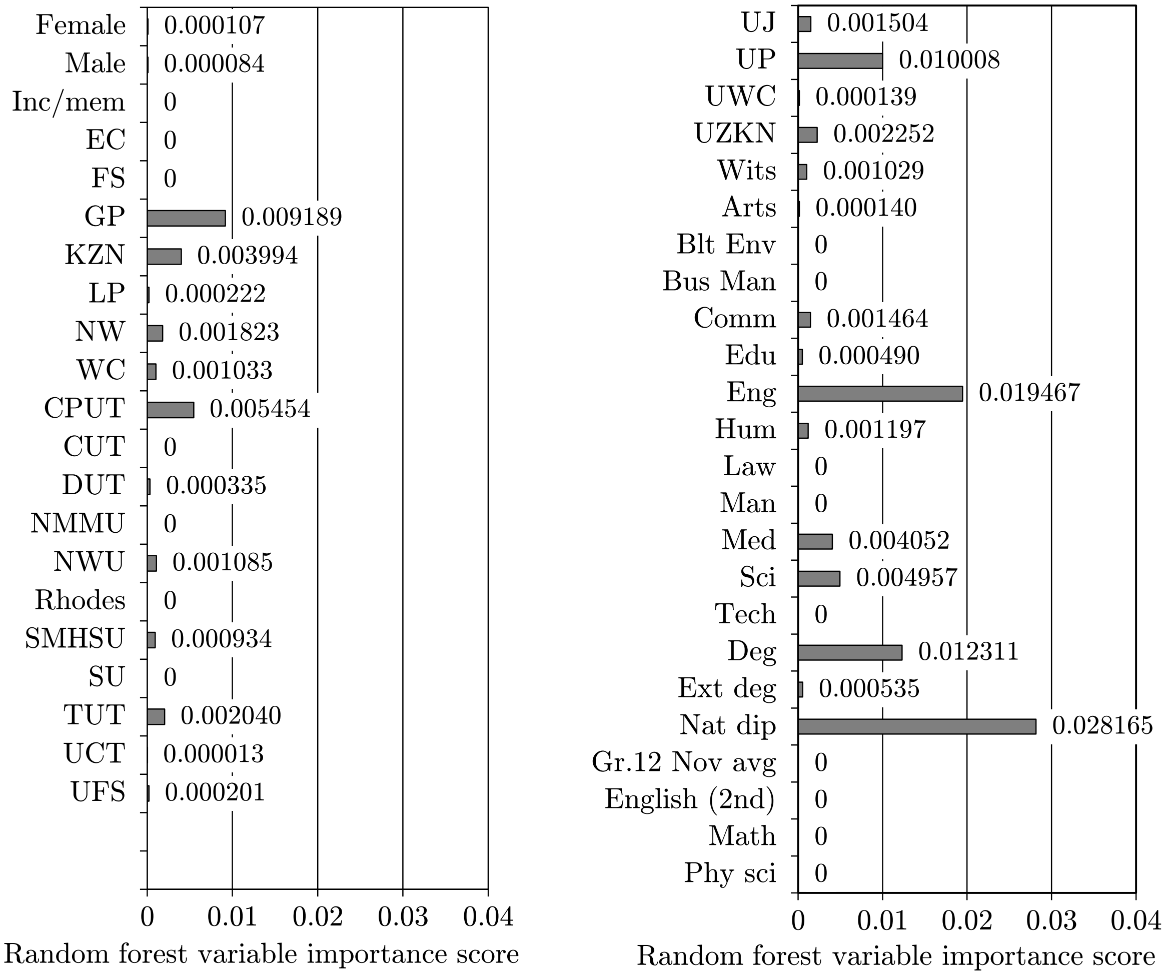


FIGURE 8.18: *Random forest variable importance scores for Combinations 33 and 34 together, including dummy variables, of Case Study B.*

Based on the above random forest variable importance analysis, it would seem that the following findings produced by the logistic regression models may be considered cautiously and preliminarily corroborated. *Study region*, *Tertiary institution*, *Qualification type*, and *Study field* appear to be able to predict the success of students at tertiary level. At the end of the logistic regression analysis, the variables *Gender*, *Grade 12 November average*, *Grade 12 Mathematics*, and *Grade 12 Physical science* were furthermore identified as potentially also being able to predict tertiary success, but now it would appear not be the case after all. *Grade 12 June average* has furthermore been shown to be potentially valuable as a tertiary success predictor, but the small sample size used in the combinations leading to this finding should be taken into consideration.

CART plots for the classification models

The third step was to evaluate the model combinations by investigating their CART plots. One plot was drawn up for each of the combinations that showed potential to predict tertiary success of students. Note that as with the random forest variable importance scores, those combinations that were combined so as to avoid multicollinearity violations again have their variables combined for the CART plots (that is, Combination 27 with Combination 28, Combination 29 with Combination 30, and Combination 33 with Combination 34).

These eighteen CART plots are shown in Figures 8.19–8.35. The key presented in Figure 8.3 may again be used to interpret nodes in these CART plots.

From Figure 8.19, containing the CART plot of *Income per family member*, it would at first glance seem that this variable lacks any real overall positive or negative trend in terms of classifying students in nodes that are either more likely to succeed or more likely to withdraw. Note, however, that the first partition of placing those students associated with a family income per member of at least R 4 116.00 in a single child node is merely a case of isolating the outliers. Apart from this, the remaining splits do not indicate any real trend, which agrees with the previous findings in which *Income per member* was identified as not having real predictive potential in terms of tertiary success.

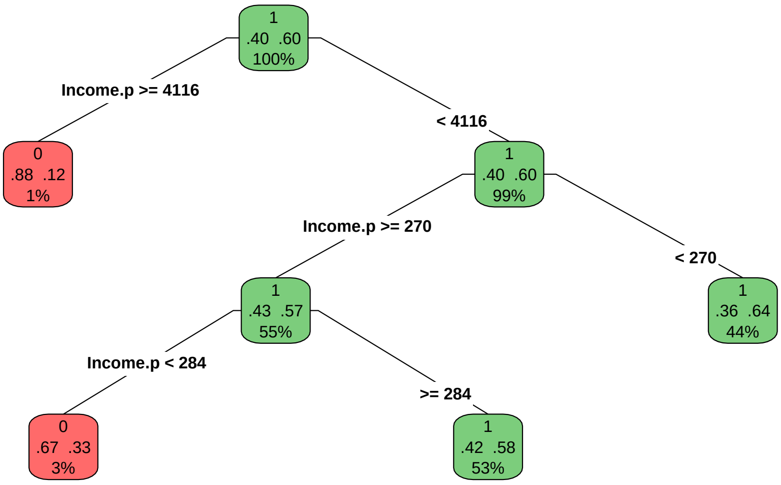


FIGURE 8.19: *CART plot for Combination 3 of Case Study B.*

The variable *Study region* is contained within the CART plot shown in Figure 8.20. Based on the importance associated with this variable up to now, one would be inclined to take the partitions of the categories of *Study region* seriously. From the three terminal nodes created, it would appear that *Gauteng* contains the observations that achieve the lowest tertiary success rate, while *KwaZulu-Natal*, *Limpopo*, and *North West*, contain the observations that achieve the highest success rate.

This agrees with the previous findings that *Gauteng* and *KwaZulu-Natal* are the most important categories in terms of tertiary success predictive power. Note that although *Gauteng* is the only category contained within its node, the proportion of observations in the node that belong to one class is still very close to 50%, which does not bode well for its tertiary success predictive ability. In contrast, the terminal node containing *KwaZulu-Natal*, *Limpopo*, and *North West*, are contained in a node where the proportion of students that belong to one class is more skewed at 74%, which indicates good predictive ability of tertiary success.

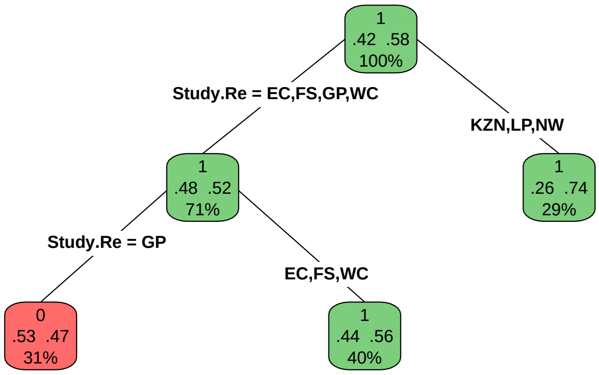


FIGURE 8.20: *CART plot for Combination 4 of Case Study B.*

The next CART plot, displayed in Figure 8.21, relates to the variable *Tertiary institution*. As also previously noted, this variable seems to have strong predictive power in terms of tertiary success. The terminal node containing *Rhodes University*, *Stellenbosch University*, *University of Pretoria*, and *University of the Witwatersrand*, contain those observations that correspond to the lowest tertiary success rate. On the other side, the terminal node containing *Cape Peninsula University of Technology*, *Durban University of Technology*, *North West University*, *Sefako Makgatho Health Sciences University*, *Tshwane University of Technology*, and *University of KwaZulu-Natal* contain the variables which correspond to the highest tertiary success rate. The remaining three terminal nodes, containing the other categories, all have observations whose proportions of tertiary success lie very close to 50%, indicating poor tertiary success predictive power.

The CART plot of Combination 6 (shown in Figure 8.22) contains the singular variable *Study field*. The two significant terminal nodes are the outer ones. The node containing *Engineering* and *Law* corresponds to a tertiary success rate of 41% for its observations. The other extremal terminal node holds *Arts*, *Business management*, *Education*, *Humanities*, *Management*, and *Medical*, with a tertiary success rate of 75% for observations contained therein. Again this variable seems to have the potential for good tertiary success predictive power.

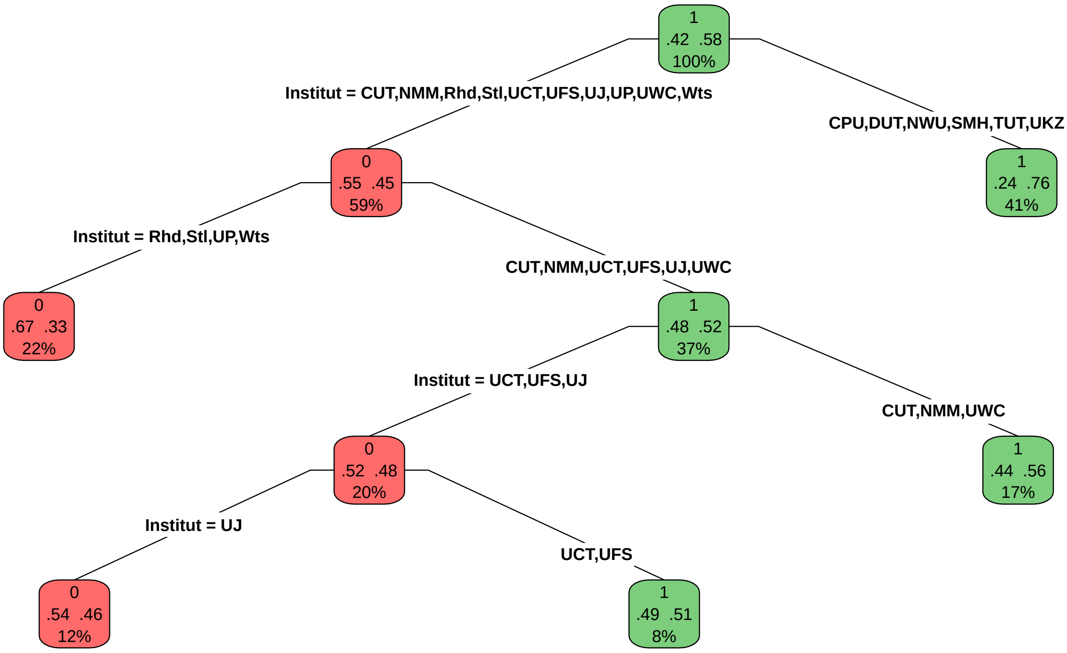


Figure 8.21: *CART plot for Combination 5 of Case Study B.*

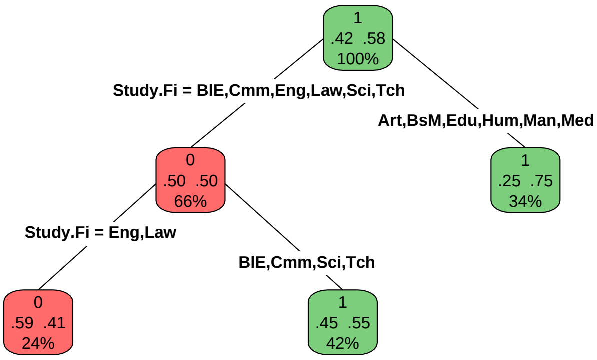


Figure 8.22: *CART plot for Combination 6 of Case Study B.*

Qualification type, another variable previously identified as important, is the only variable present in the CART plot of Figure 8.23. Clearly, the most valuable category in respect of predicting tertiary success is *National diploma*, as its observations achieve a historical tertiary success rate of 74%, as seen in the right-most terminal node. From the left-most terminal node it would appear that the observations corresponding to *Extended degree* achieve a success rate of 45%, which is the lowest of the three categories. The final category, *Degree*, basically seems to hold no predictive power in terms of tertiary success as its observations' success rate is 50%.

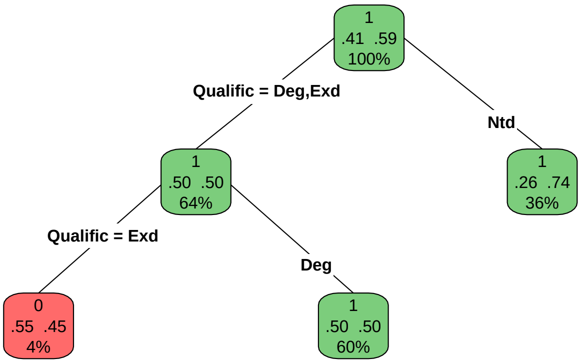


FIGURE 8.23: *CART plot for Combination 7 of Case Study B.*

Combination 8 (whose CART plot is shown in Figure 8.24), contains *Grade 11 November average*. It generally appears that a lower *Grade 11 November average* mark leads to a higher success rate, thus confirming the previous findings, and there are a few exceptions. These exceptions inhibit a strong negative trend, in agreement with the logistic regression model finding that the CI widths are narrow for *Grade 11 November average*, but that its $\hat{\beta}$-coefficient is very close to zero.

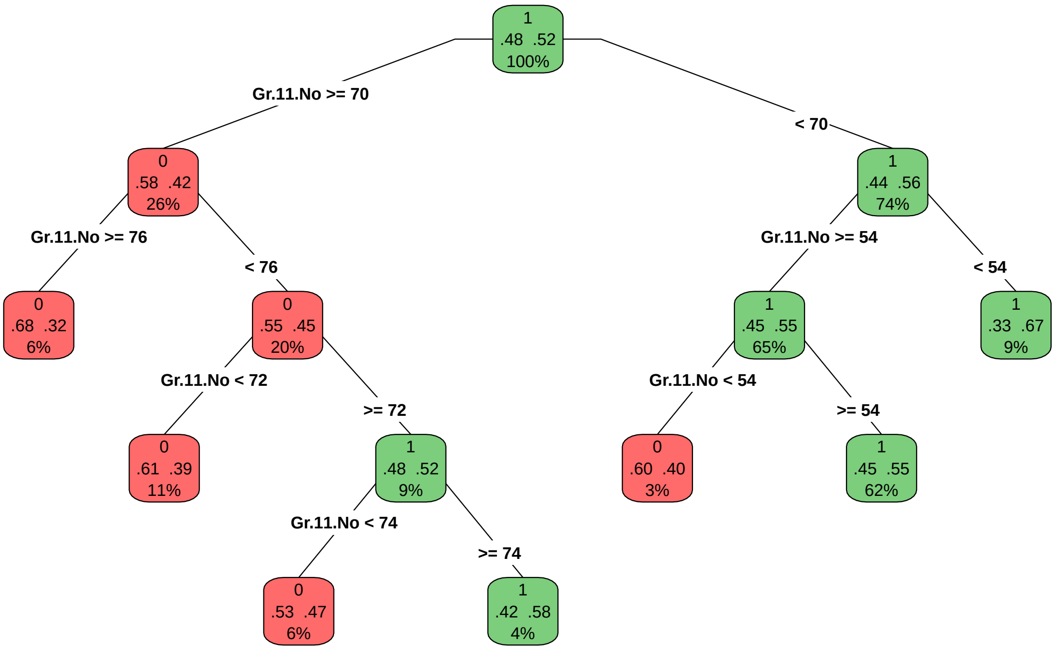


FIGURE 8.24: *CART plot for Combination 8 of Case Study B.*

As with *Grade 11 November average*, it would appear for *Grade 12 June average*, whose CART plot is shown in Figure 8.25, that a higher mark for this variable leads to a higher probability of failure at tertiary level. Once again, although this is the clear general trend, there also appears to be many exceptions judging from some of the terminal nodes, which reduces the overall strength of this negative relationship between the variable and tertiary success. Also notice how both *Grade 11 November average* and *Grade 12 June average* have their first partition at 70% — this will be discussed again later.

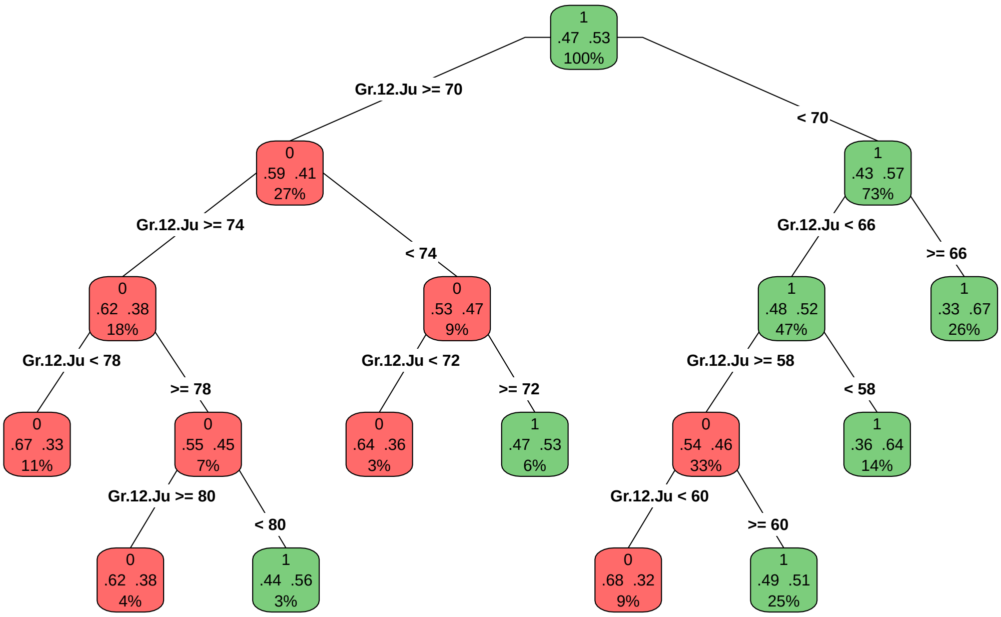


FIGURE 8.25: *CART plot for Combination 9 of Case Study B.*

The CART plot of the last of the three high school averages, *Grade 12 November average*, is shown in Figure 8.26. Although the probabilities of success of the observations corresponding to the two outer terminal nodes are very small and very large, respectively, note that each only contains 1% of the sample and hence not too many conclusions should be drawn from them. The only other potential trend in the figure relates to the partition of the left-most child node of the root node. It would appear that once again a higher *Grade 12 November average* mark leads to a higher tertiary withdrawal rate. It also appears, however, that the slope of this trend is not very steep.

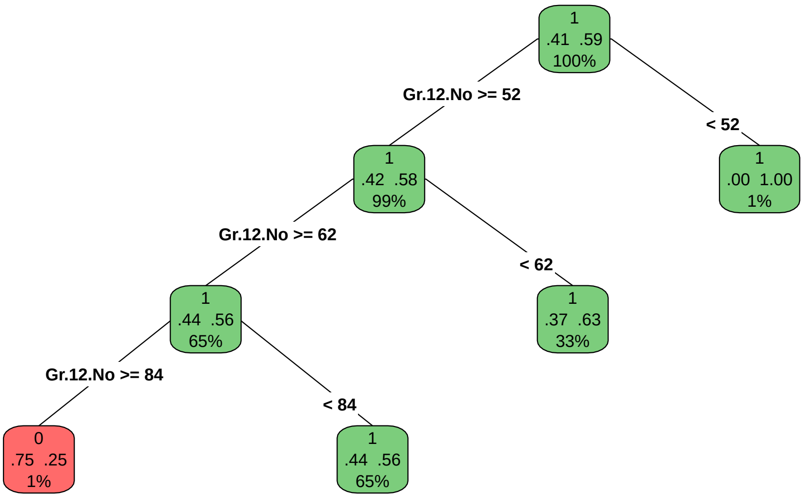


FIGURE 8.26: *CART plot for Combination 10 of Case Study B.*

From the CART plot of the first of the four high school subjects, *Grade 12 Accounting* shown in Figure 8.27, it does not appear that any significant trend is noticeable in terms of tertiary success prediction, as partitioning of the sample into those of both higher and lower *Grade 12 Accounting* marks lead to higher success rates.

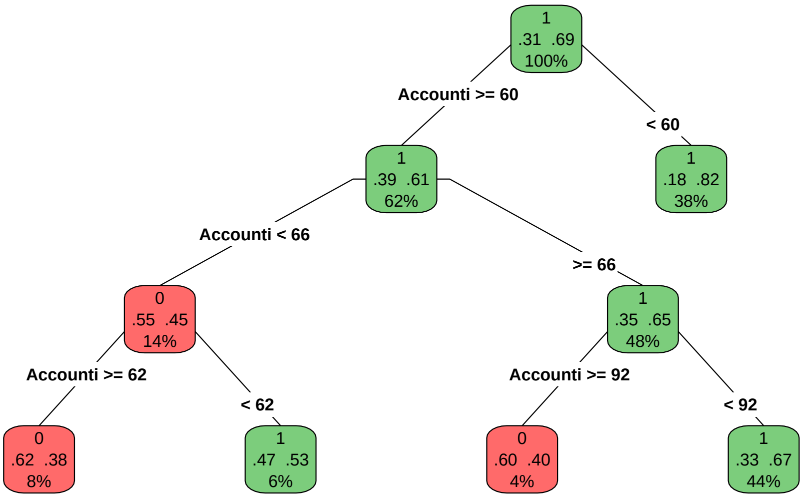


FIGURE 8.27: *CART plot for Combination 11 of Case Study B.*

The CART plot of *Grade 12 English (2nd)* is shown in Figure 8.28. The left-most and central terminal nodes are ignored in the interpretation of the results, because the left-most terminal node contains only 1% of the sample, thus concerning too small a percentage of the sample to draw conclusions from, and the make-up of the central terminal node is not much different from its parent node. By focussing on the remaining three nodes, *i.e.* only the root node and its two child nodes, it may again be seen that an increase in a high school subject, this time *Grade 12 English (2nd)*, leads to a lower tertiary success rate, even if only slightly.

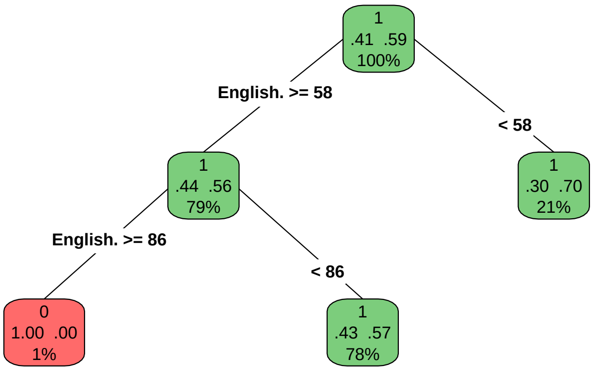


FIGURE 8.28: *CART plot for Combination 17 of Case Study B.*

The CART plot of *Grade 12 Mathematics* is shown in Figure 8.29. Just based on the first two partitions of the tree, the all-to-familiar trend is observed that an increasing high school mark produces a lower probability of tertiary success.

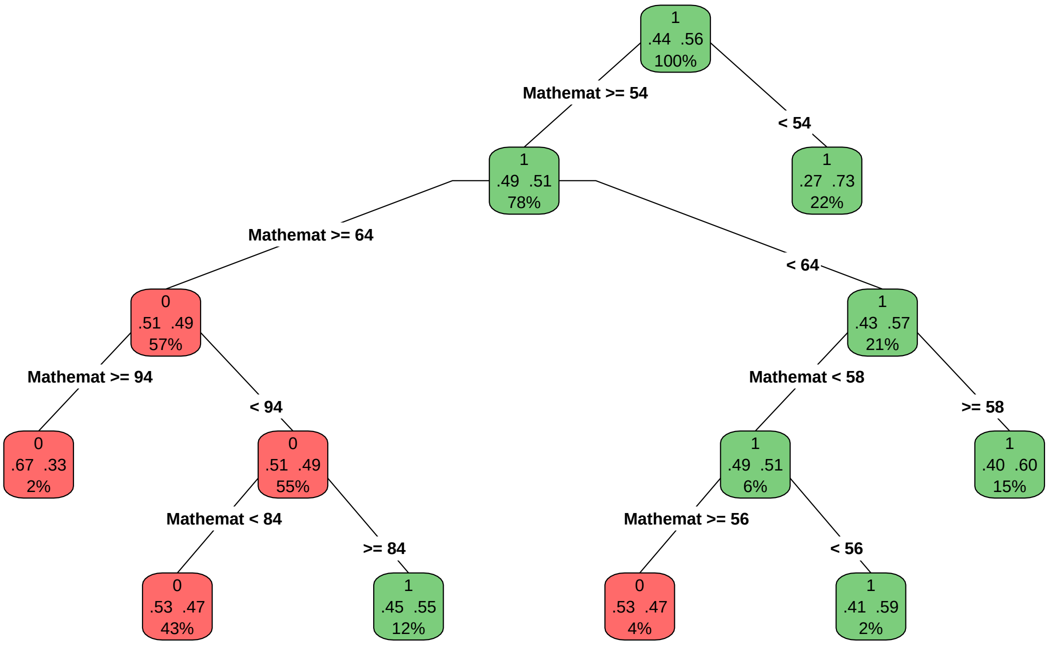


FIGURE 8.29: *CART plot for Combination 23 of Case Study B.*

The last of the four highs school subjects, *Grade 12 Physical science*, may be further investigated by considering its CART plot shown in Figure 8.30. From the top three partitions it should be clear that, just as with *Grade 12 Mathematics*, those students who achieve higher *Grade 12 Physical science* marks are less likely to succeed at tertiary level.

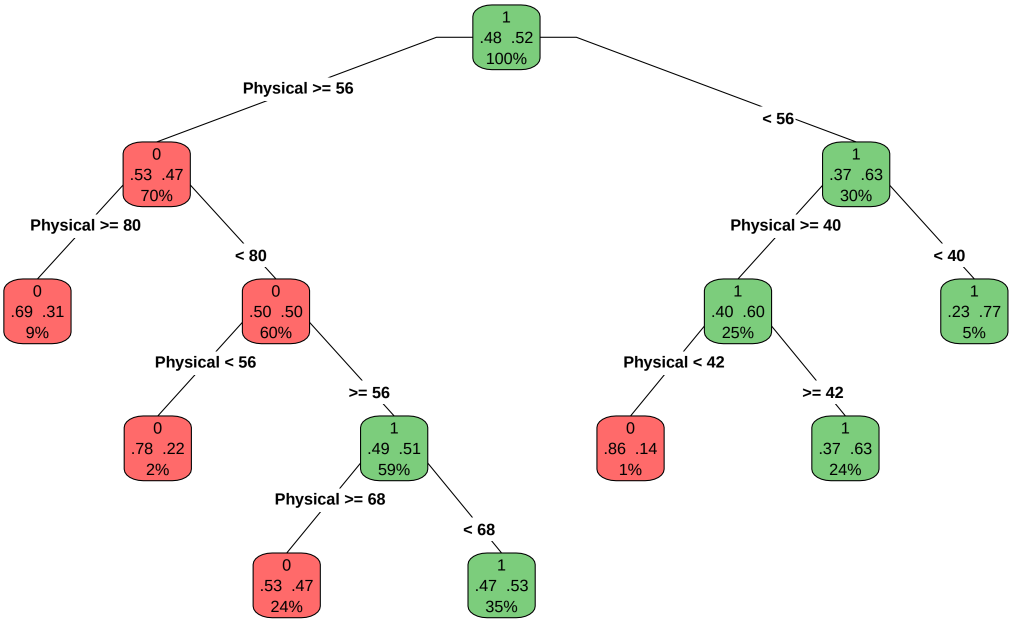


FIGURE 8.30: *CART plot for Combination 25 of Case Study B.*

Based on the CART analyses of the high school averages and subject marks, the theory that increasing high school marks lead to a higher withdrawal rate seems to be reinforced at this point. Despite this, it is disappointing that many of these high school variables which initially showed signs of predictive potential based on the logistic regression and random forests analyses seem much less so now.

The CART plot of the combined variables of Combinations 27 and 28 are shown in Figure 8.31, which contains the variables *Gender*, *Race*, *Income per family member*, *Study region*, *Tertiary institution*, *Study field*, and *Qualification type*.

As expected, the partitioning points for each variable are identical or nearly identical to the partitions found in the previous CART plots that contain the variables separately which are also contained in Combinations 27 and 28. It is, however, interesting to note the order in which the variables are partitioned in the tree. *Tertiary institution* is partitioned first, followed by partitioning of *Study field*, then *Study region* and finally again *Tertiary institution*. The *Income per family member* partition is ignored as it would seem more a case of removing an outlier, as are the remaining partitions which deal with small subsets of the sample.

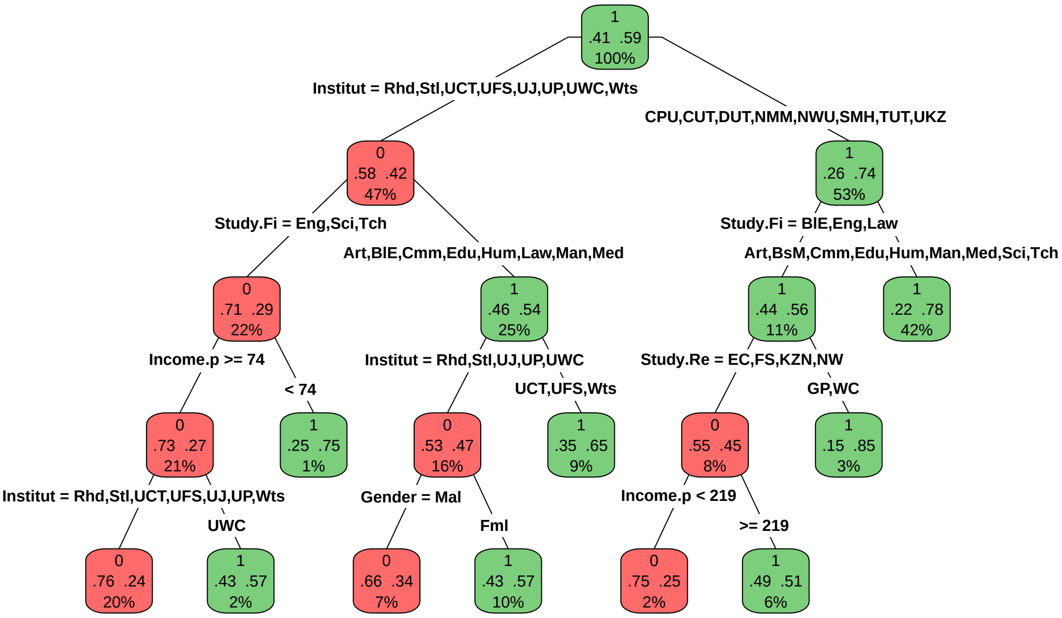


Figure 8.31: *CART plot for Combinations 27 and 28 together of Case Study B.*

Combinations 29 and 30 (see Figure 8.32), are identical to Combinations 27 and 28, with the addition of *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average*. Although the sample size of this CART plot is reduced significantly to 254 observations from 863, the top two partitions occur on the variables *Tertiary institution* and *Study field*. The third-level partition involves *Study region* again and the newly added variable *Grade 12 November average*.

The results of the next CART plot, shown in Figure 8.33, should not be afforded too much consideration as it was not able to produce a single important variable according to the random forest variable importance scores. Despite this, and the fact that a small sample is used, it should come as no surprise that once again it appears that an increase in any of the high school marks lead to lower tertiary success rates.

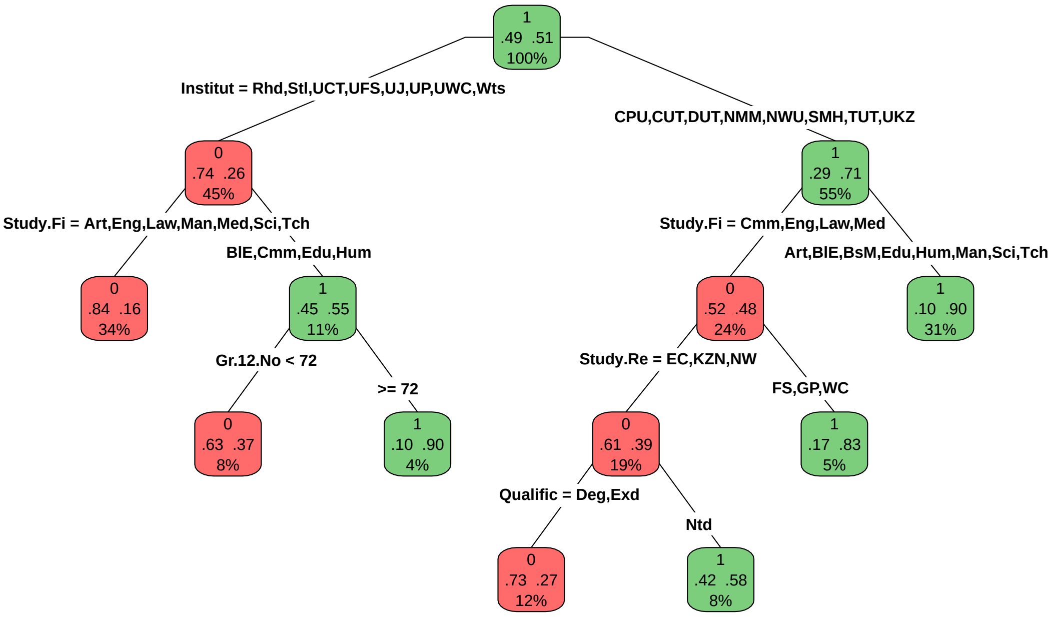


FIGURE 8.32: *CART plot for Combinations 29 and 30 together of Case Study B.*

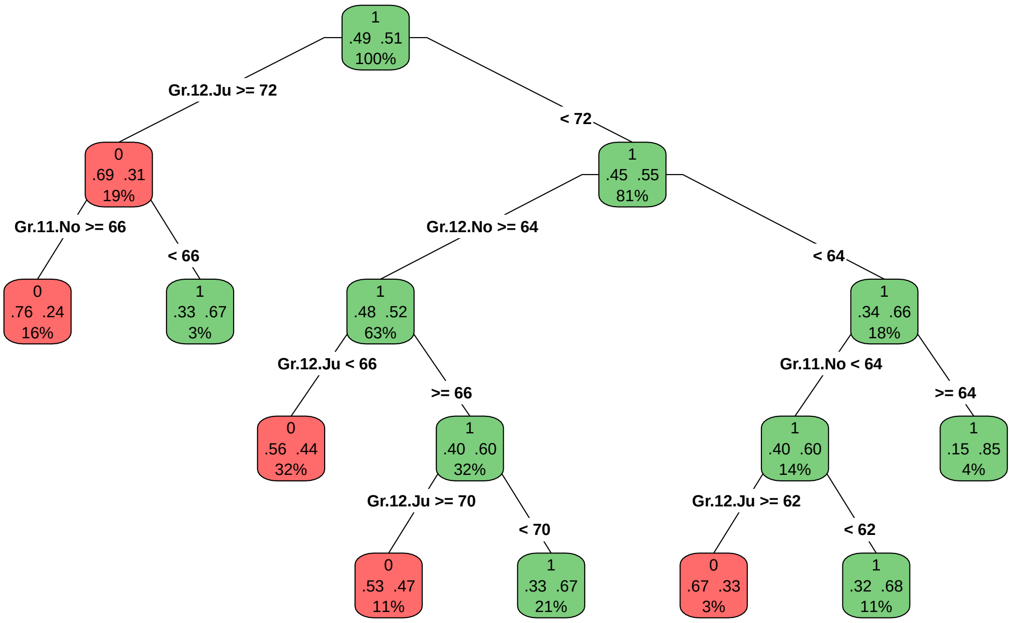


FIGURE 8.33: *CART plot for Combination 31 of Case Study B.*

The CART plot of Combination 32, containing *Grade 11 November average*, *Grade 12 June average*, *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science* is shown in Figure 8.34. Only three splits occur and they involve *Grade 12 June average* (which agrees with the random forest variable importance score where *Grade 12 June average* was the only notable variable of the six), *Grade 12 Physical science*, and *Grade 12 English (2nd)*, in that order. Once again it is clear than an increase in any of these marks leads to a higher withdrawal rate at tertiary level.

The combined variables of Combinations 33 and 34 are included in the CART plot of Figure 8.35. These combinations are identical to Combinations 27 and 28, with the exclusion of *Race* and addition of *Grade 12 November average*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and

Grade 12 Physical science. Despite the addition of these variables, the partitions of the top nodes involve similar variables as on which branching occurred in the CART plot of Combinations 27 and 28.

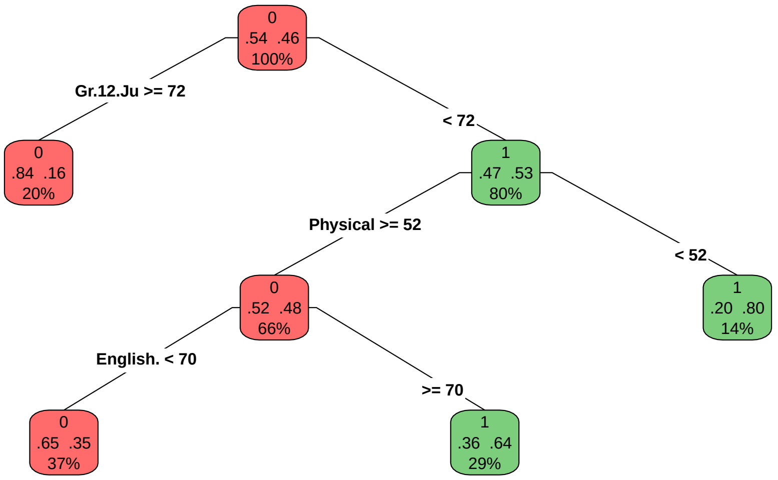


FIGURE 8.34: *CART plot for Combination 32 of Case Study B.*

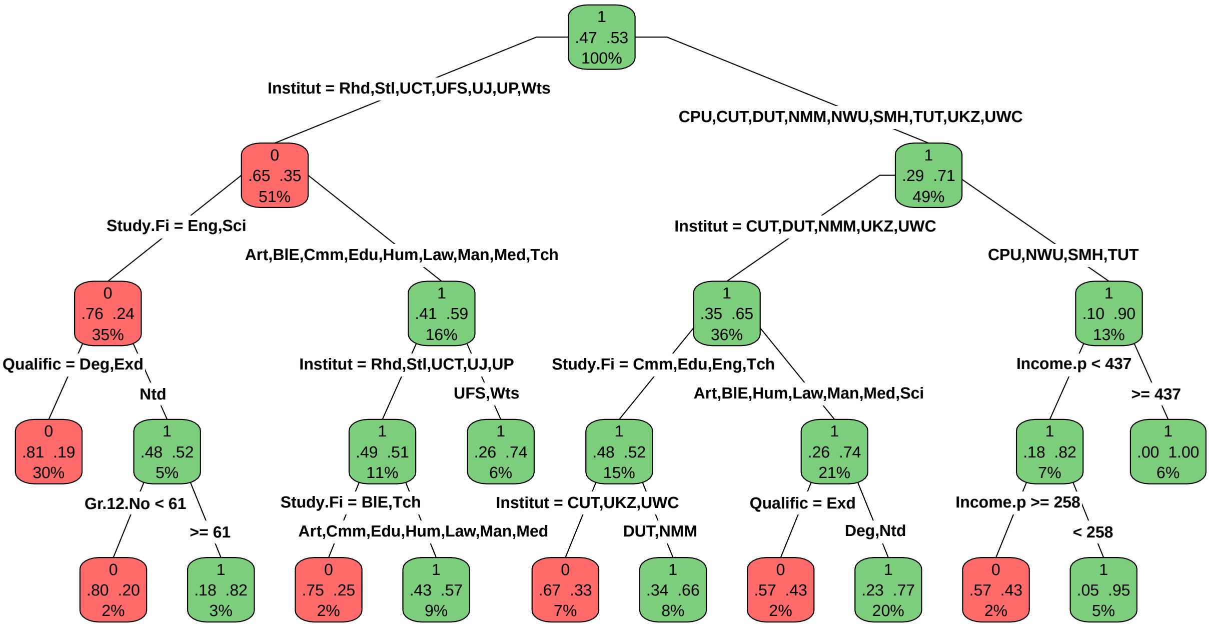


FIGURE 8.35: *CART plot for Combinations 33 and 34 combined of Case Study B.*

Based only on the logistic regression model analysis and random forest variable importance score analysis, it appeared that the variables *Study region*, *Tertiary institution*, *Study field*, and *Qualification type* are of importance in terms of tertiary success predictive power. At this stage, however, it would appear that the top three variables in respect of predicting tertiary success are *Tertiary institution*, *Study region*, and *Study field.* This does not mean that *Qualification type* is not of value in this respect, but rather that in comparison to the others it is not as valuable in terms of predicting tertiary success of prospective students.

Although the high school variables performed disappointingly in the above respect in all three analyses performed so far, it can be confirmed that despite the weak trend that is present, there is negative correlation between these variables and the tertiary success of students.

Bar plots of the variables of interest

The final step in evaluating the individual variables is to draw up bar plots indicating the percentages of students who graduated successfully as functions of specific independent variables. These plots for all twenty six independent variables of Table 8.2, except variables 12–16, 18, 20–22, 24, and 26, are shown in Figures 8.36–8.48.

The corresponding figures for high school mark independent variables that have not shown much predictive potential may be found in Appendix C. These are the independent variables of *Grade 12 Afrikaans home language*, *Grade 12 Afrikaans second language*, *Grade 12 Agricultural science*, *Business economics*, *Grade 12 Economics*, *Grade 12 Geography*, *Grade 12 History*, *Grade 12 isiXhosa home language*, *Grade 12 isiZulu home language*, *Grade 12 Life science*, *Grade 12 Mathematics literacy*, and *Grade 12 Setswana home language.*

Upon consideration of Figure 8.36 it is evident why *Race* has shown no indication of being a good predictor variable — its two categories *African* and *Coloured* have near identical success rates.

The percentage of *Male* and *Female* applicants who graduated successfully, shown in Figure 8.37, validates the previous finding that *Gender* is a good predictor of the tertiary success of a prospective student. While *Male* by itself does not hold predictive power (it corresponds to a success rate of 50%), *Female* is both significantly different to *Male* and reasonably far away from 50%. This plot also corroborates the negative $\hat{\beta}$-coefficient obtained for *Male* when *Female* was the reference class in the logistic regression analysis.

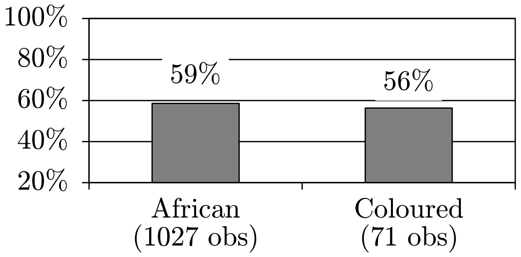


FIGURE 8.36: *Percentages of students who graduated successfully as a function of race in the sample of Case Study B.*

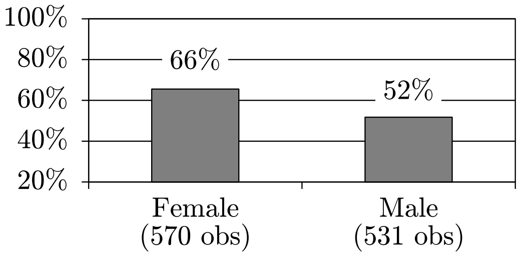


FIGURE 8.37: *Percentages of students who graduated successfully as a function of gender in the sample of Case Study B.*

Figure 8.38 contains a bar plot of the percentages of students who graduated successfully as a function of *Study region.* Upon investigation of the figure it should be clear why *Gauteng* and *KwaZulu-Natal* are the categories exhibiting the most predictive promise in terms of tertiary success, as they are the ones containing the lowest and highest proportions of students who were successful in respect of their tertiary studies, respectively. Overall the fluctuation of the bars in Figure 8.38 is an indication of a good predictor (in the case of qualitative variables).

The same type of bar plot for the sixteen categories of *Tertiary institution* is shown in Figure 8.39. The overall plot indicates clear fluctuation between the bars, which supports the previous finding that *Tertiary institution* is one of, if not the most, valuable variables with respect to predicting tertiary success of students. It is expected that the highest and lowest bars should correspond to the best predictive categories. The low bars corresponding to *Rhodes University* and *Stellenbosch University* should explain why they consistently received such high p-values in the logistic regression model analyses — they are visibly not much different in the figure from the *University of Pretoria*, which was the reference class.

The bar plot of the next variable, *Study field*, is shown in Figure 8.40. Like the previous two variables there exist considerable fluctuations between the categories, indicating that overall it should be a good predictor variable for tertiary success.

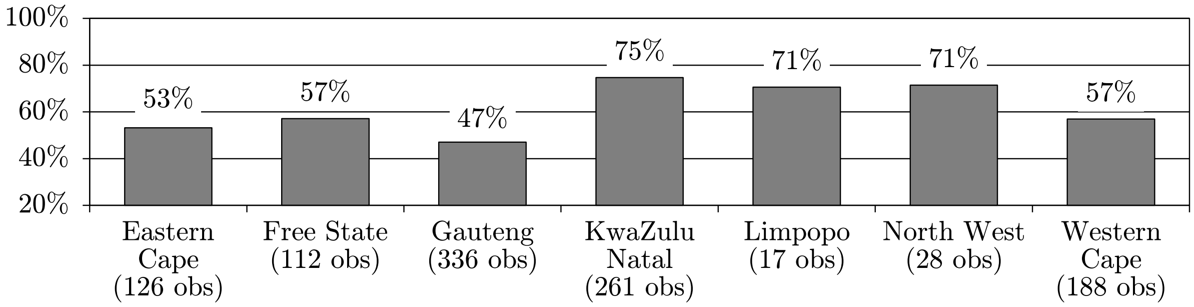


FIGURE 8.38: *Percentages of students who graduated successfully as a function of study region in the sample of Case Study B.*

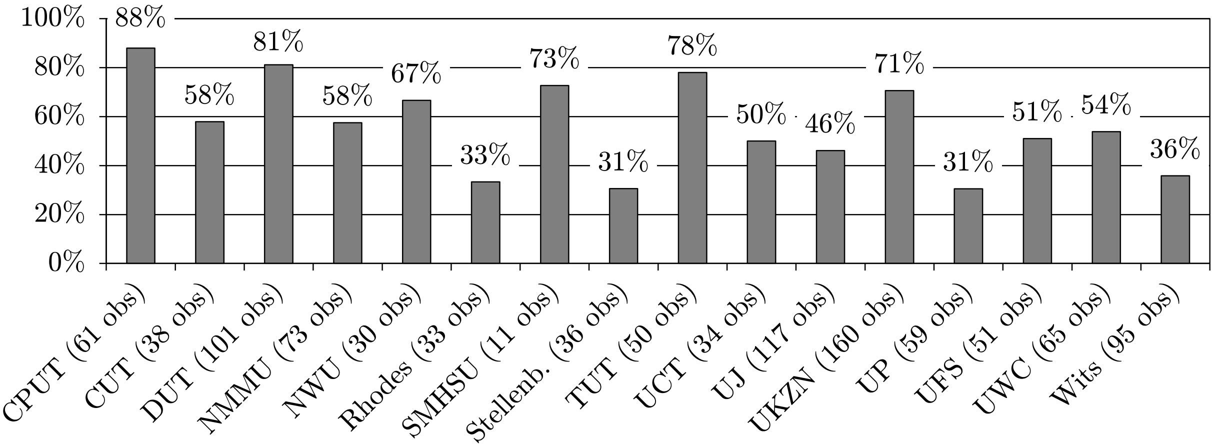


FIGURE 8.39: *Percentages of students who graduated successfully as a function of tertiary institution in the sample of Case Study B.*

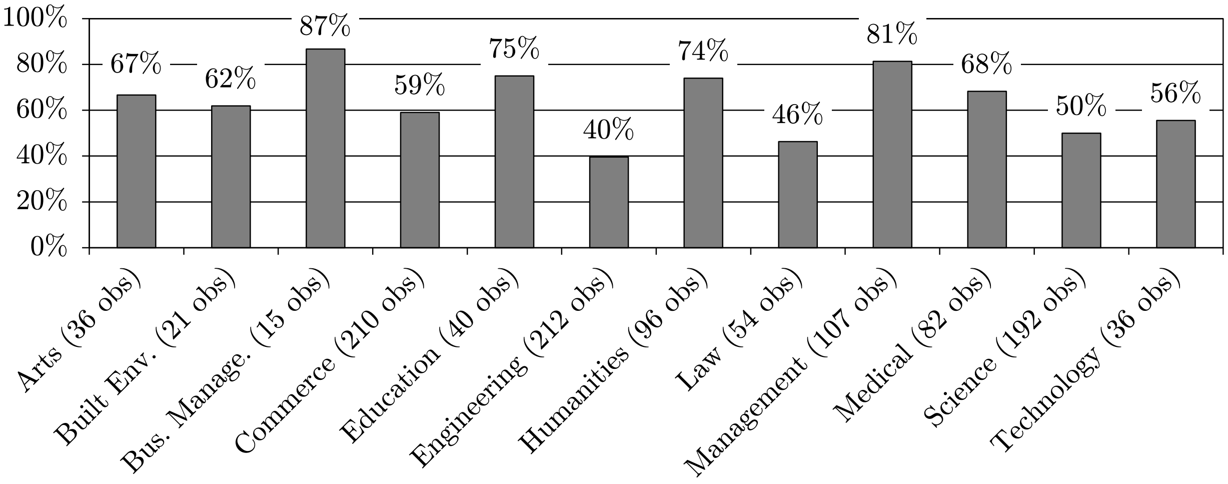


FIGURE 8.40: *Percentages of students who graduated successfully as a function of study field in the sample of Case Study B.*

Figure 8.41 contains the same type of plot for the three categories of *Qualification type*. A possible reason for why *Qualification type* might have been considered important in the previous analyses, and especially its category *National diploma*, but may have been considered overall less important than the previous three variables, is that both *Degree* and *Extended degree*, containing large potions of the observations, lie very close to 50%. Despite this, the high bar for *National diploma* still makes this a valuable variable.

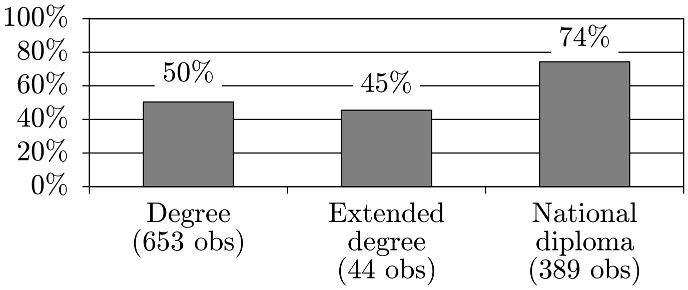


FIGURE 8.41: *Percentages of students who graduated successfully as a function of qualification type in the sample of Case Study B.*

The bar plot for the three high school variables *Grade 11 November average*, *Grade 12 June average*, and *Grade 12 November average* are shown in Figures 8.42–8.44. Ignoring the first bar of each of the three plots (due to the small number of observations), it is clear that there exists a slight, but definite, downward slope in each of the plots as a function of increasing marks. This confirms the finding in the logistic regression and CART analyses that those students who who achieve better high school marks are less likely to succeed at tertiary level. The fact that the slopes are also only slightly negative agree with the small $\hat{\beta}$-coefficients assigned to these variables in the logistic regression analysis.

Now consider specifically the variables of *Grade 11 November average* and *Grade 12 June average* and their CART plots in Figures 8.24–8.25. Recall that the first partition in both plots occurred at 70%, which is the separation point between an average mark of a C and a B. This may be confirmed by investigating the bar plots of the two variables displayed in Figures 8.42–8.43. Notice how the bars remain reasonably constant over the 50%–69% bins, but then drop off towards the 70%–79% bin. Although more data are required to state this with confidence, a possible reason for this phenomenon in these two variables, which is not present in *Grade 12 November average*, is that these two marks generally play a more significant role than *Grade 12 November average* in the admission of students to tertiary institutions. This is because prospective students are provisionally accepted based on these two marks after which they only need to achieve a minimum required mark for their *Grade 12 November average*. Based on this observation, it may be argued that students who achieve higher marks for *Grade 11 November average* and *Grade 12 June average* will go on to attend more challenging tertiary institutions and study fields, and thus experience a higher drop-out rate.

The final four bar plots contain the tertiary success rates as a function of marks for the high school subjects *Grade 12 Accounting*, *Grade 12 English (2nd)*, *Grade 12 Mathematics*, and *Grade 12 Physical science*, and are shown in Figures 8.45–8.48. Less so for *Grade 12 Accounting* and *Grade 12 English (2nd)*, but for the other two (as for the previous three variables), there exists a slight negative slope as marks increase. This, of course, confirms the previous findings of negative $\hat{\beta}$-coefficients and smaller than one ORs during the logistic regression analysis. It may hinder *Grade 12 Mathematics* and *Grade 12 Physical science* from achieving higher importance scores that there exist four and three bars in their plots, respectively, which lie very close to success rates of 50%.

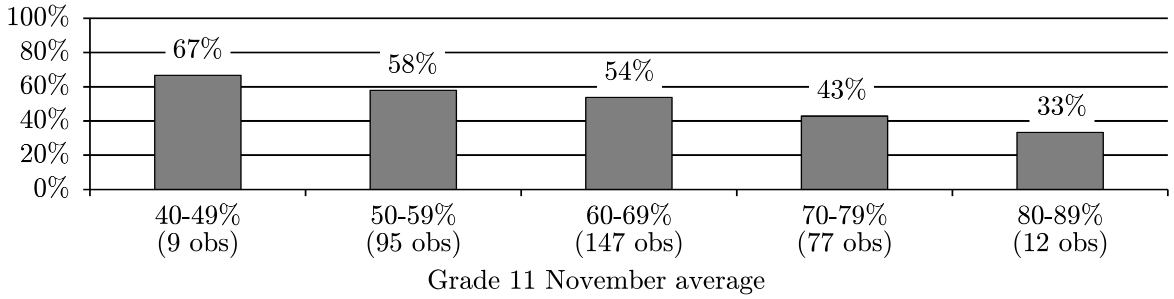


FIGURE 8.42: *Percentages of students who graduated successfully as a function of Grade 11 November average in the sample of Case Study B.*

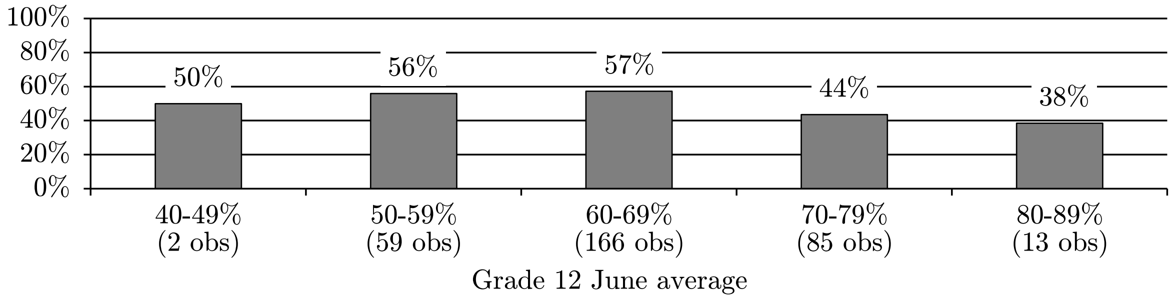


FIGURE 8.43: *Percentages of students who graduated successfully as a function of Grade 12 June average in the sample of Case Study B.*

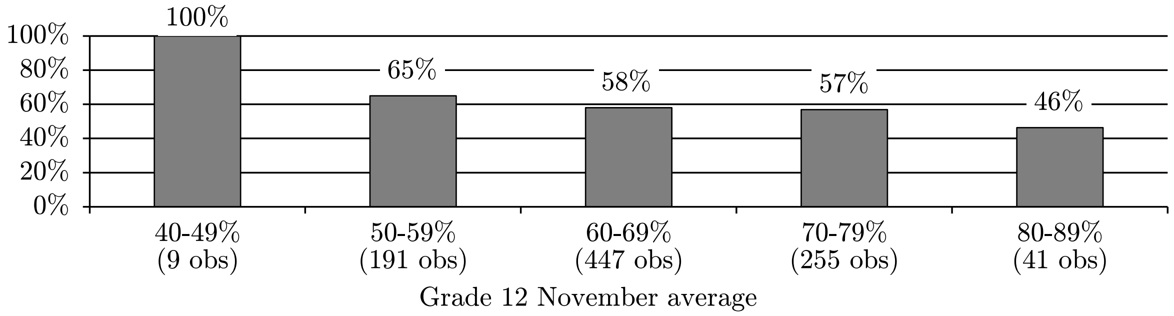


FIGURE 8.44: *Percentages of students who graduated successfully as a function of Grade 12 November average in the sample of Case Study B.*

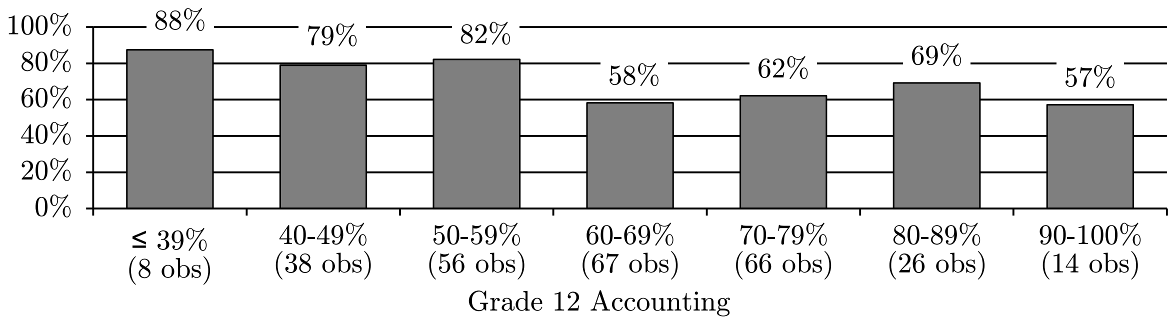


FIGURE 8.45: *Percentages of students who graduated successfully as a function of Grade 12 Accounting mark in the sample of Case Study B.*

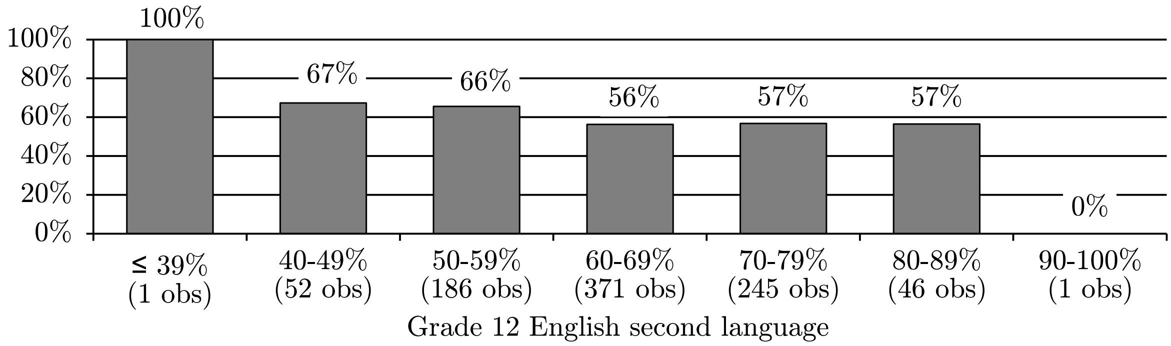


Figure 8.46: *Percentages of students who graduated successfully as a function of Grade 12 English (2nd) mark in the sample of Case Study B.*

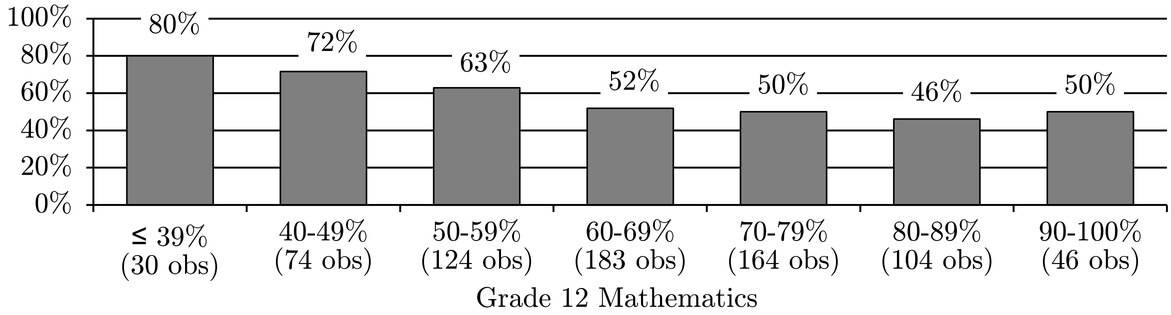


Figure 8.47: *Percentages of students who graduated successfully as a function of Grade 12 Mathematics mark in the sample of Case Study B.*

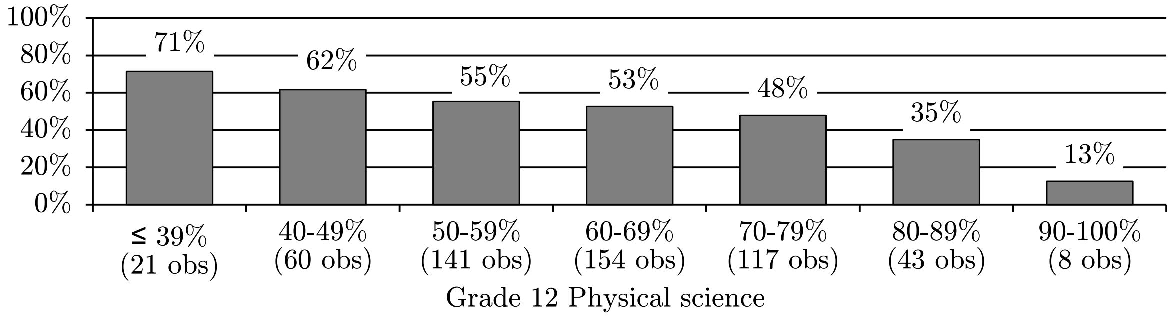


Figure 8.48: *Percentages of students who graduated successfully as a function of Grade 12 Physical science mark in the sample of Case Study B.*

At this point it should be clear that the same rather counter-intuitive trend is again present that as students' marks for specific high school averages and high school subjects increase, their probability of success at tertiary level decreases. As mentioned before, a suspicion, which will further be elaborated on later, is that those students who achieve higher marks at school are admitted to pursue more challenging qualifications, in more challenging study fields, and at more challenging tertiary institutions. Despite the above evidence it would, of course, be foolish to offer bursaries to those students who achieve the lowest marks at school. The remainder of this subsection is devoted to investigating the source of this peculiar trend further so that more sensible recommendations may be made to the main industry partner.

Consider Table 8.30, which is concerned with the *Grade 12 November averages* of students *versus* their *Tertiary institutions*. The table is partitioned into a left-hand and a right-hand side. The left-hand side indicates the proportion of students within each percentage class (*e.g.* 50%–59%) of *Grade 12 November average* that attended a specific tertiary institution. The percentages within each column of the table therefore add up to 100%. These column entries are also shaded.

The largest percentage (highest attendance per tertiary institution), is shaded light grey, while the lowest attendance is shaded dark grey. The rows, in this case tertiary institutions, are ordered from those who have the smallest graduation percentage at the top to those who have the highest graduation percentage at the bottom. For example, of those students who obtained a Grade 12 November average between 80%–89%, 10% attended Stellenbosch University.

The right-hand side of the table indicates the proportion of students within each percentage class of *Grade 12 November average* who graduated successfully at a specific tertiary institution. The shading for the right-hand side occurs across the entire area, where those blocks with the lowest graduation rate are shaded dark grey, while those blocks with the highest graduation rate are shaded light grey. Following on the previous example, of those students who obtained a Grade 12 November average between 80%–89% and attended Stellenbosch University, none therefore graduated successfully. Those right-hand side cells which are not applicable due to the corresponding 0% attendance rates on the left-hand side of the table are crossed out. The total number of observations within each of the percentage classes is listed in the final row on both sides of the table.

Tables of the above form are presented for *Grade 12 November average versus Tertiary institution* (see Table 8.30), *Study field* (see Table 8.31), *Qualification type* (see Table 8.32), and *Study region* (see Table 8.33). Similarly, tables of this kind were also drawn up for *Grade 11 November average*, *Grade 12 Mathematics*, and *Grade 12 Physical science*, *versus Tertiary institution*, *Study field*, *Qualification type*, and *Study region*. These tables may be found in Appendix C.

When considering the left-hand side of Table 8.30, a clear trend is visible. Although exceptions are present, it would seem that students who obtained higher Grade 12 November average marks tend to attend those tertiary institutions that have a lower graduation rate, indicated by the light grey shading and high percentages in the top right, while students who obtained lower Grade 12 November average marks tend to attend those tertiary institutions that have a higher graduation rate, similarly indicated by the light grey shading and high percentages in the bottom left. The obvious exception to this trend is the University of KwaZulu-Natal, which is considered a tertiary institution with a higher graduation rate, but is attended by a large proportion of students who achieved high Grade 12 November average marks.

The reason for this may be geographical. If the better performing students (*i.e.* those who achieved high Grade 12 November average marks) desired to study within their local provinces, those who originate from the Western Cape, Eastern Cape, Gauteng, and Free State, all have an opportunity to attend a tertiary institution that is ranked in the top half of those with a lower graduation rate and which are thus more challenging and potentially more prestigious. On the other hand, the higher performing students who originate from KwaZulu-Natal and desire to study at a local tertiary institution are left with the University of KwaZulu-Natal as their primary option.

By investigating the right-hand side of Table 8.30, a second trend is visible, namely that those students in the top half and top right-hand area are shaded darker, indicating a lower graduation rate, although they achieved higher Grade 12 November average marks. Similarly, those students in the bottom and bottom-left achieve much higher graduation rates, although they have lower Grade 12 November average marks.

These two trends are again evident in Tables 8.31–8.33, concerned with *Study field*, *Qualification type*, and *Study region*. First, students who achieve higher high school marks tend to gravitate towards the more challenging study fields, qualification types, and study regions. Secondly, students who attend the more challenging tertiary institutions have a lower graduation rate, and *vice versa*, although they have superior high school marks.

TABLE 8.30: *Proportions of students within each percentage class of Grade 12 November marks who attended a specific tertiary institution and graduated at a specific tertiary institution, respectively.*

	Proportions of students within each % class of Gr.12 Nov marks that:									
	attended a tertiary institution					graduated at a tertiary institution				
	40-49	50-59	60-69	70-79	80-89	40-49	50-59	60-69	70-79	80-89
SU	0%	2%	3%	4%	10%	☒	25%	33%	18%	0%
UP	0%	5%	5%	7%	5%	☒	60%	23%	12%	0%
Rhodes	0%	2%	4%	4%	0%	☒	0%	39%	33%	☒
Wits	0%	2%	7%	18%	32%	☒	50%	37%	35%	38%
UJ	0%	10%	14%	10%	10%	☒	37%	37%	76%	25%
UCT	0%	0%	3%	5%	12%	☒	☒	58%	46%	60%
UFS	0%	4%	4%	5%	5%	☒	38%	61%	50%	100%
UWC	0%	6%	8%	4%	0%	☒	36%	56%	67%	☒
CUT	0%	5%	4%	3%	0%	☒	80%	50%	57%	☒
NMMU	0%	5%	8%	4%	0%	☒	90%	59%	67%	☒
NWU	0%	4%	3%	2%	2%	☒	63%	77%	50%	100%
UKZN	20%	9%	14%	22%	20%	100%	82%	70%	69%	63%
SMHSU	0%	1%	0%	0%	2%	☒	100%	☒	☒	100%
TUT	0%	6%	4%	3%	2%	☒	73%	82%	100%	100%
DUT	40%	15%	10%	8%	0%	100%	71%	82%	90%	☒
CPUT	40%	8%	7%	2%	0%	100%	75%	87%	100%	☒
# obs	5	161	428	254	41	5	101	247	144	19

TABLE 8.31: *Proportions of students within each percentage class of Grade 12 November marks who studied within a specific field of study and graduated in a specific field of study, respectively.*

	Proportions of students within each % class of Gr.12 Nov marks that:									
	studied in a specific study field					graduated in a specific a study field				
	40-49	50-59	60-69	70-79	80-89	40-49	50-59	60-69	70-79	80-89
Eng	0%	14%	19%	22%	34%	☒	44%	43%	36%	14%
Law	0%	3%	6%	4%	2%	☒	40%	31%	55%	0%
Sci	0%	13%	19%	22%	24%	☒	58%	47%	50%	60%
Tech	0%	2%	3%	4%	5%	☒	75%	53%	55%	50%
Comm	11%	15%	18%	22%	17%	100%	54%	61%	63%	43%
Blt Env	0%	1%	2%	2%	0%	☒	100%	44%	67%	☒
Arts	0%	6%	3%	2%	2%	☒	55%	92%	60%	100%
Med	11%	4%	8%	8%	2%	100%	75%	69%	71%	100%
Hum	11%	9%	8%	9%	10%	100%	76%	70%	83%	100%
Edu	0%	5%	6%	0%	2%	☒	78%	81%	☒	100%
Man	56%	26%	7%	4%	0%	100%	80%	81%	100%	☒
Bus Man	11%	4%	1%	0%	0%	100%	71%	100%	☒	☒
# obs	9	191	447	255	41	9	124	259	145	19

Table 8.32: *Proportions of students within each percentage class of Grade 12 November marks who pursued a specific qualification type and obtained a specific qualification type, respectively.*

	Proportions of students within each % class of Gr.12 Nov marks that:									
	pursued a qualification type					obtained a qualification type				
	40-49	50-59	60-69	70-79	80-89	40-49	50-59	60-69	70-79	80-89
Ext Deg	0%	5%	5%	3%	5%	☒	33%	64%	43%	0%
Degree	11%	42%	57%	78%	93%	100%	59%	48%	49%	47%
Nat Dip	89%	53%	38%	20%	2%	100%	72%	73%	88%	100%
# obs	9	188	441	255	41	9	122	256	145	19

Table 8.33: *Proportions of students within each percentage class of Grade 12 November marks who studied in a specific study region and graduated in a specific study region, respectively.*

	Proportions of students within each % class of Gr.12 Nov marks that:									
	studied in a specific study region					graduated in a specific study region				
	40-49	50-59	60-69	70-79	80-89	40-49	50-59	60-69	70-79	80-89
GP	29%	27%	31%	38%	54%	100%	57%	42%	47%	41%
EC	14%	8%	12%	7%	0%	100%	73%	53%	50%	☒
FS	0%	14%	10%	9%	5%	☒	64%	60%	55%	100%
WC	14%	17%	20%	15%	22%	100%	55%	63%	49%	33%
LP	0%	5%	1%	0%	0%	☒	70%	33%	100%	☒
NW	0%	4%	2%	2%	0%	☒	63%	91%	50%	☒
KZN	43%	25%	24%	30%	20%	100%	76%	75%	75%	63%
# obs	7	183	441	253	41	7	118	257	143	19

8.3.8 Bootstrapped predictive accuracy assessment

The twenty two variable combinations that have shown predictive potential are put to the test in this section by training the five base models, implemented in the DSS of Chapter 6, in respect of a bootstrapped sample, with replacement, and testing the results against the remaining OOB observations for 100 iterations. The average predictive performance of the weighted predictions produced by the DSS for each of the twenty two variable combinations may be seen in Table 8.34.

The predictive accuracy produced is, on average, around 60% for combinations containing only one variable, and increases to around 70% for combinations involving multiple variables. The exception is Combination 11, containing only *Grade 12 Accounting*, which achieves a high predictive accuracy of tertiary success compared to the other combinations. This may partly be due to the slight negative trend visible in the frequency bar plot of Figure 8.45. It may also, however, to some extent be due to skewness in the data. The term *skewness* here refers to how the proportion of one outcome class exceeds that of the other. The proportion of observations that belong to the largest class of the samples of Combination 11 (*Grade 12 Accounting*) is 69%.

Recall, from Table 8.3, that the samples of these two combinations only contained 267 and 111 observations, respectively. These sample sizes are small in comparison to those of many of the other combinations, and hence it is possible that students who took these subjects reside in a specific segment of the samples which makes prediction easier. For example, consider again the root nodes of the CART plots in Figures 8.27–8.28. From these figures it should be clear that the samples of the two combinations are skewer than those of the samples of the other combinations.

It is also possible to compare the performance of some of the pairs of combinations created to avoid multicollinearity, that is Combinations 27 and 28, and Combinations 33 and 34. In each case, the first combination of each pair, that is Combinations 27 and 30, achieve slightly higher predictive accuracy than the second combinations. Since each pair of combinations is identical, except that the first combination includes *Study region* whereas the second does not, but rather includes *Tertiary institution*, it seems that the inclusion of *Study region*, as opposed to *Tertiary institution*, in a logistic regression model increases its predictive ability. This is slightly unexpected, since the predictive accuracy of the combination containing only *Study region* (Combination 4) is lower than that of Combination 5 which contains only *Tertiary institution*.

It is also important to keep the skewness of the individual samples in mind when considering these predictive accuracy results. By referring back to the root nodes of the CART plots of the variable combinations, it is possible to correlate the skewness of the samples for each combination with its predictive performance. Combinations 3, 7, 10, 11, and 17 exhibit nearly no difference and thus achieve very poor predictability. Combinations 1, 4, 6, and 23 furthermore exhibit small differences of around 3%. The combinations that exhibit large differences of around 10% are Combinations 5, 8, 9, 19, 25, 27, 28, 31, and 32. Finally, Combinations 29, 30, 33, and 34 exhibit very large differences of around 19%. In addition, it should be noted that Combinations 29, 31, and 32 have relatively small samples compared to the other combinations. Based on the above it would seem that Combinations 30, 33, and 34 achieve the highest predictability based on a stable sample.

Table 8.34: *Average weighted prediction accuracy of selected combinations tested on 100 bootstrapped samples of the data of Case Study B.*

Combination number	Average weighted prediction accuracy	Combination number	Average weighted prediction accuracy
1	60.46%	23	60.23%
3	60.82%	25	60.12%
4	61.17%	27	67.19%
5	64.33%	28	66.97%
6	61.71%	29	71.83%
7	59.89%	30	69.24%
8	60.33%	31	62.02%
9	61.31%	32	64.48%
10	60.47%	33	68.86%
11	69.18%	34	67.77%
17	60.88%		

8.3.9 Report back of results

It should be stressed that the data sample used in this case study has serious limitations, including small samples for some variable combinations, uneven prior class probabilities, missing values which induce the creation of multiple combinations and subsequently multicollinearity among many of the variables. Despite these limitations, the best attempt was made at analysing the data responsibly.

Due to the above limitations and the possibility that the recommendations contained in this section may directly influence the allocation of funding to students and potentially provide opportunities that would otherwise not have existed, it would be irresponsible to report back very specific results confidently. For these reasons the data were purposefully analysed in a more general manner, searching for trends rather than focussing on specific values, such as ORs.

Based on the analyses of the data according to the methodologies of logistic regression, random forest variable importance, CART plots and frequency bar plots, discussed in depth above, the variables *Study region*, *Tertiary institution*, *Study field*, and *Qualification type*, may possibly be considered to have predictive power in terms of the tertiary success of students. Other variables that might be able to aid in this prediction, but do not rank nearly as important, are the high school academic variables *Grade 11 November average*, *Grade 12 June average*, *Grade 12 November average*, *Grade 12 Mathematics*, and *Grade 12 Physical science.*

The finding that high school marks play a relatively small role in predicting tertiary success compared to the other qualitative variables is, of course, somewhat surprising. Linked to this is perhaps the most interesting finding, namely that negative relationships exist between school marks and tertiary success. This peculiar trend was investigated thoroughly, and it was shown in Tables 8.30–8.33 that a possible reason for this trend may be that higher achieving high school students tend to gravitate towards the more challenging *Study regions*, *Tertiary institutions*, *Study fields*, and *Qualification types*, where they are much more likely to withdraw as opposed to their peers.

The trend of higher achieving high school students experiencing a larger drop out is obviously very concerning and a difficult problem to solve. The answer is, of course, not to offer bursaries to the worse performing high school students. It would appear that a conundrum is at play here. Consider the three factors of cost of a tertiary bursary (*i.e.* to the main industry partner), the quality of a qualification (*i.e.* how prestigious the qualification is and how easily students will find work based on such a qualification), and the probability of tertiary success (*i.e.* the overall success rate at tertiary students in the past). Generally speaking, it may be assumed that the cost of a bursary is higher for universities as opposed to universities of technology. It may also, generally speaking, be assumed that the quality of a qualification is positively related to the cost of a tertiary bursary. The third factor that the probability of tertiary success is negatively related to is the cost of a tertiary bursary and thus also the quality of education. Another consideration which makes this conundrum even more challenging is the 'running cost' of a student. Based on cost alone, it may be said that it would have been a smaller expense to the main industry partner if a student dropped out in their first year, as an even better student who also did not make it and dropped out in their third year. The answer to how to approach this conundrum is not straightforward, but bringing these findings and trends to the attention of the main industry partner is potentially helpful the first step.

Although many supposedly correlation relationships may have appeared to show up in the above analyses, it is important to consider the possible reasons for correlational relationships. The reader is referred back to §3.2, where a discussion was conducted on possible reasons for correlation. The two most important ones identified there were measurement error and possible confounders.

While it was implicitly assumed above that all data were accurately collected, it is possible that measurement errors may have occurred, whereby applicants provide incorrect high school academic information and incorrect family income information. Although unlikely, due to quality checks employed by the industry partner, this occurrence should not be ruled out completely.

Another important possible reason for correlation relationships is confounders. For example, factors such as the difficulty of qualifications, the academic standards of the various tertiary institutions and the difficulty of the various study fields are confounders to *Qualification type*, *Tertiary institution*, and *Study field*, as they will be correlated to both the independent variables, respectively, and the dependent variable. Other examples of possible confounders on which no data were collected, but may well be very informative, is psychological assessment results, family environment, and other support structures.

There are many other unrecorded factors that may also potentially shed light on student success trends. This does not, however, imply that the results of this study are inferior, as it is perhaps the best that can be done in view of the data available. These findings may furthermore be used to justify more intense and careful data collection by the organisations.

8.4 Case Study C

The data for and aim of Case Study C were presented in §7.5 and §7.6, respectively.

As in previous case study, the roadmap for assessing variable importance, as outlined in §8.2, is applied here in a sequential manner.

8.4.1 Selection of a single dependent variable

The variable *Academic status* is again selected as the binary dependent variable, as specified in §7.6.

8.4.2 Identification of independent variables and reference classes

The eleven currently remaining independent variables may be seen in Table 7.25.

Of the 96 entries in the *High school institution* field, 17 are blank and among the remaining 79 entries there are 71 high schools present. Due to this large ratio of high school institutions to the number of non-blank entries, this variable is not considered for further analysis.

In the same fashion, the *Qualification* field contains 52 different categories while the *Specialisation* field contains 24 different categories. Although smaller than the number of entries in the *High school institution* field, this equates to an average of about only two and four students per category of the fields, respectively. This makes both prediction and variable assessment impractical when using these variables.

The *Admission point score*, although potentially a good predictor of tertiary success, is calculated slightly differently by each tertiary institution. For example, the University of the Witwatersrand includes life orientation in their *Admission point score* calculation, while the University of KwaZulu-Natal does not include life orientation. This makes comparing the *Admission point score* across different tertiary institutions by combining it in a single logistic regression model impossible. The *Admission point score* field is thus also removed from the data.

After these exclusions, only seven independent variables remain, as listed in Table 8.35.

As explained in §8.2, qualitative variables are converted to multiple levels using the method of dummy variables. As was also explained, the reference class or baseline class for each variable is chosen as the category containing the smallest proportion of students who were successful at their tertiary studies. The selected reference classes for each remaining qualitative independent variable is also shown in Table 8.35.

8.4.3 Identification of valid combinations of remaining variables

No multicollinearity was found between any combination of independent variables, and hence the only remaining constraint is the sample size. It was decided to consider the combinations of fields listed in Table 8.36. As may be seen, there is one combination for each of the seven single independent variables, one containing only the high school academic marks as it is thought this might produce interesting results, and one containing all the applicable independent variables.

Of these, Combinations 1, 4, 5, 6, and 7 satisfy the minimum sample size requirements for logistic regression. All of the remaining combinations are nevertheless cautiously analysed, as explained in §8.2.

Table 8.35: *Independent variables and reference classes for the sample data of Case Study C.*

	Field number	Data field name	Reference class	Proportion of reference class outcome labelled 0
Personal	1	Gender	Male	33.3%
	2	Race	African	21.7%
Academic administration	3	Tertiary institution	Wits	13.5%
	4	Study field	Engineering	21.7%
High school academics	5	Grade 12 Mathematics	N.A.	N.A.
	6	Grade 12 Physical science	N.A.	N.A.
	7	Grade 12 English	N.A.	N.A.

Available and required observations for additional combinations of Case Study B

Table 8.36: *Number of available and required observations for each combination of Case Study C.*

Combination number	Fields in combination	Number of fields	Number of variables and categories	Number of observations: in each combination	required by (3.78)
1	1	1	2	96	62
2	2	1	4	95	123
3	3	1	6	91	202
4	4	1	2	96	62
5	5	1	1	96	31
6	6	1	1	96	31
7	7	1	1	96	31
8	5,6,7	3	3	96	93
9	1,2,3,4,5,6,7	7	17	90	567

8.4.4 Identification of outliers and influential values

The numbers of outliers and influential observations identified for each combination are shown in Table 8.37. Note that, except for Combination 9, no outliers are identified and thus none are deleted from Combinations 1–8. Due to the shortage of observations in the sample, it was, however, decided that the single influential value identified in Combination 9 would not be deleted.

8.4.5 Validation of the remaining assumptions

Once the decision was made as to whether or not to delete the outliers and influential observations, the remaining two assumptions of independence of residuals and linearity of the logit were considered for each of the nine combinations, the results of which are shown in Table 8.37. For

TABLE 8.37: *Number of outliers and influential observations identified, and assessment of the remaining assumptions for each combination of Case Study C.*

			Actions taken:		Assumption of:		
Combination number	Num of outliers	Num of inf obs	Outliers deleted?	Inf obs deleted?	Independence of residuals	Multi-collinearity	Linearity of the logit
1	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
2	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
3	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
4	0	0	N.A.	N.A.	Satisfied	N.A.	N.A.
5	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
6	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
7	0	0	N.A.	N.A.	Satisfied	N.A.	Satisfied
8	0	0	N.A.	N.A.	Satisfied	Satisfied	Satisfied
9	0	1	N.A.	No	Satisfied	Satisfied	Satisfied

all nine observations, all three assumptions (including multicollinearity discussed above) were satisfied and so it was unnecessary to exclude of any of the combinations from the analysis.

8.4.6 Evaluation of the overall model fits and addition of new combinations

Logistic regression models for each of the combinations in Table 8.36 were assessed against a null model so as to obtain the log likelihood p-values shown in Table 8.38.

Although Combinations 1, 5, 6, 7, and 8 do not exhibit a low log likelihood p-value, the total number of combinations of this case study are a mere ten, and so these combinations are nevertheless analysed in the following steps.

TABLE 8.38: *Log likelihood p-values for each combination of Case Study C.*

Combination number	Log likelihood p-value	Combination number	Log likelihood p-value	Combination number	Log likelihood p-value
1	0.804352	**4**	**0.000525**	7	0.470521
2	**0.002966**	5	0.353447	8	0.219612
3	**0.083036**	6	0.215738	**9**	**0.000043**

Based on these overall model fit results, those combinations consisting of only a single independent variable and having the smallest log likelihood p-values (Combinations 2, 3, and 4) were combined and added as Combination 10. The same assessments as were performed for the other nine combinations were also performed for this new combination, the results of which are shown in Tables 8.39–8.41. As for some of the other combinations, the sample size (see Tables 8.39) is not adequate for Combination 10, but the remaining assumptions (see Table 8.40) are satisfied. As expected, the log likelihood p-value produced for the combination (see Table 8.41) is lower than that of any of the other nine combinations at 0.0013%.

TABLE 8.39: *Number of available and required observations for the additional variable combination of Case Study C.*

				Number of observations:	
Combination number	Fields in combination	Number of fields	Number of variables and categories	in each combination	required by (3.78)
10	2,3,4	3	12	90	375

Table 8.40: *Number of outliers and influential observations identified for the additional variable combination of Case Study C.*

			Actions taken:		Assumption of:		
Combination number	Num of outliers	Num of inf obs	Outliers deleted?	Inf obs deleted?	Independence of residuals	Multi-collinearity	Linearity of the logit
10	0	0	N.A.	N.A.	Satisfied	Satisfied	Satisfied

Table 8.41: *Log likelihood p-values for the additional variable combination of Case Study C.*

Combination number	Log likelihood p-value
10	**0.000013**

8.4.7 Evaluation of the individual independent variables

The ten combinations described in §8.4.3–8.4.6 are now analysed so as to evaluate the performance of the individual independent variables in the corresponding logistic regression models.

Their p-values, $\hat{\beta}$-coefficients, ORs, $\hat{\beta}$-CIs, and OR CIs

The p-values, $\hat{\beta}$-coefficients, $\hat{\beta}$-CIs, and OR CIs for the various independent variables and their respective categories in each of the ten combinations are shown in Tables 8.42–8.51.

From Table 8.42 it may be seen that for Combination 1 the Category *Gender* of the *Female* variable is not significantly different to that of the reference class *Male* due to the large p-value of 80.46%. As then expected, the $\hat{\beta}$-CI contains the value zero. The OR was therefore not investigated further. Similarly, the individual p-values produced for Combinations 5 (see Table 8.46), 6 (see Table 8.47), 7 (see Table 8.48), and 8 (see Table 8.49) do not indicate statistical significance. This should come as no surprise since these five combinations performed the worst according to the log likelihood fits of Table 8.38. These early results also echo those of the previous case study (in §8.3), namely that those variables traditionally expected to be most influential when predicting tertiary success, such as the school marks, are not as significant as other variables, such as *Study field* and *Tertiary institution.*

Another trend, which was also previously observed, is that an increase in the performance of school marks leads to a higher tertiary drop out rate. This finding may be justified by considering Table 8.49 of Combination 8. The only p-value less than 10% is that of *Grade 12 Physical science* at 8.07%. The $\hat{\beta}$-coefficient is negative, although only slightly so and the OR is just less than one. By considering the CI, one may be excused to suspect that the finding of a significant p-value is unsupported, since the 95% CI contains zero, but this is because the p-value is above 5%, while it is reasonable to expect that with a 90% CI this would not be the case. Although counter-intuitive, this observation is very similar to what was witnessed in the previous case study, although not as significant, and calls for further exploration.

Next, consider the remaining three combinations that contain only one independent variable, namely Combinations 2 (see Table 8.43), 3 (see Table 8.44), and 4 (see Table 8.45). For all three combinations, the small log likelihood p-values listed in Table 8.38 are found, as expected. It is also not unexpected that the $\hat{\beta}$-coefficients are positive and that the ORs greater than one in all these tables since the reference classes for each was chosen as the category with the lowest proportion of students who graduated at tertiary level.

TABLE 8.42: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 1 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.016344	−0.6931	−1.2589	−0.1274	1.1316	0.5000	0.2840	0.8804	0.5965
Female	0.804557	−0.1092	−0.9741	0.7557	1.7298	0.8966	0.3775	2.1291	1.7516

For Combination 2 (see Table 8.43) it is shown that *Indian* and *White* applicants are statistically significant from *African* applicants. They are shown to be more likely to be successful at tertiary level than their *African* counterparts due to the positive $\hat{\beta}$-coefficients and large ORs. The OR may be interpreted as follows: the odds of an *Indian* student being successful at tertiary studies over the odds of an *African* student being successful at tertiary studies is 6.18. Notice, however, that the precision of these estimates is not very high due to the relatively large CI widths, but this may be due to the small sample size. The effect of larger CI widths due to the small sample size is present for all the combinations of this case study. Note also that no pronouncement can be made about whether or not *Indian* and *White* applicants are statistically different from one another.

TABLE 8.43: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 2 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000016	−1.2622	−1.8355	−0.6890	1.1464	0.2830	0.1595	0.5021	0.3425
Indian	0.008438	1.8219	0.4662	3.1775	2.7113	6.1833	1.5940	23.9863	22.3923
White	0.004883	2.1095	0.6406	3.5785	2.9379	8.2444	1.8976	35.8197	33.9222
Coloured	0.533554	0.5691	−1.2225	2.3606	3.5831	1.7667	0.2945	10.5978	10.3033

For Combination 3 (see Table 8.44), the University of Pretoria is shown to be statistically the most different from the University of the Witwatersrand, with a large OR of 7.31. Again its CI is rather wide. Excluding the University of Johannesburg, the other tertiary institutions have p-values of around 10%, meaning that they are not considered statistically indistinguishable from the University of the Witwatersrand just yet.

In Combination 4 (see Table 8.45), the *Study field* category of *Science* is shown to be highly statistically different from *Engineering*. Also, due to the positive $\hat{\beta}$-coefficients and relatively large OR it may be concluded that those studying *Science*-related qualifications are more likely to graduate than those studying *Engineering*-related qualifications. Although smaller than those of the above combinations, its CI widths are, however, still relatively large.

TABLE 8.44: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 3 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000113	−1.8563	−2.7988	−0.9138	1.8850	0.1563	0.0609	0.4010	0.3401
SU	0.123870	1.3455	−0.3683	3.0593	3.4276	3.8400	0.6919	21.3121	20.6202
UCT	0.100723	1.2967	−0.2517	2.8451	3.0968	3.6571	0.7775	17.2026	16.4251
UP	0.004854	1.9898	0.6052	3.3745	2.7693	7.3143	1.8316	29.2095	27.3780
UJ	0.240310	1.1632	−0.7784	3.1047	3.8830	3.2000	0.4592	22.3015	21.8423
UKZN	0.084985	1.2685	−0.1749	2.7119	2.8868	3.5556	0.8395	15.0583	14.2188

TABLE 8.45: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 4 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000012	−1.2809	−1.8530	−0.7089	1.1441	0.2778	0.1568	0.4922	0.3354
Science	0.000700	1.6556	0.6983	2.6130	1.9147	5.2364	2.0103	13.6398	11.6296

TABLE 8.46: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 5 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.221783	−2.9210	−7.6068	1.7648	9.3716	0.0539	0.0005	5.8402	5.8397
Grade 12 Math	0.358034	0.0258	−0.0292	0.0807	0.1099	1.0261	0.9712	1.0841	0.1128

TABLE 8.47: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 6 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.404811	1.6069	−2.1736	5.3873	7.5610	4.9871	0.1138	218.6228	218.5091
Grade 12 Phy sci	0.223340	−0.0313	−0.0817	0.0191	0.1008	0.9692	0.9215	1.0193	0.0977

TABLE 8.48: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 7 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.272363	−2.1057	−5.8656	1.6543	7.5198	0.1218	0.0028	5.2292	5.2263
Grade 12 English	0.472330	0.0184	−0.0317	0.0684	0.1001	1.0185	0.9688	1.0708	0.1020

TABLE 8.49: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 8 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.440442	−2.3427	−8.2945	3.6092	11.9037	0.0961	0.0002	36.9366	36.9363
Grade 12 Math	0.183759	0.0429	−0.0204	0.1062	0.1265	1.0438	0.9798	1.1120	0.1322
Grade 12 Phy sci	0.080682	−0.0498	−0.1057	0.0061	0.1118	0.9514	0.8997	1.0061	0.1064
Grade 12 English	0.384011	0.0229	−0.0286	0.0744	0.1031	1.0232	0.9718	1.0773	0.1055

Consider Combination 9 (see Table 8.50) next, which contains all the available independent variables. Those categories of the variables that were statistically significant in the previous combinations are again so. That is, the *University of Pretoria* is statistically different from the *University of the Witwatersrand*, *Indian* and *White* are statistically different from *African*, and *Science* is statistically different from *Engineering*, while controlling for each of the other variables, respectively. As before, all have positive $\hat{\beta}$-coefficients and ORs larger than one. The CI widths are the largest yet, due to the large number of independent variables present in this combination.

TABLE 8.50: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 9 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.159944	−6.7642	−16.1984	2.6701	18.8685	0.0012	0.0000	14.4408	14.4408
Grade 12 Math	0.734140	0.0146	−0.0695	0.0987	0.1682	1.0147	0.9328	1.1037	0.1709
Grade 12 Phys sci	0.305892	−0.0419	−0.1220	0.0383	0.1603	0.9590	0.8851	1.0390	0.1539
Grade 12 English	0.110096	0.0709	−0.0161	0.1578	0.1739	1.0734	0.9841	1.1709	0.1869
SU	0.517098	0.9160	−1.8554	3.6875	5.5429	2.4994	0.1564	39.9454	39.7890
UCT	0.737804	−0.4796	−3.2877	2.3285	5.6162	0.6190	0.0373	10.2621	10.2247
UP	0.022163	2.1621	0.3096	4.0145	3.7049	8.6890	1.3629	55.3951	54.0322
UJ	0.272030	1.5922	−1.2489	4.4333	5.6822	4.9145	0.2868	84.2061	83.9193
UKZN	0.704898	0.6153	−2.5691	3.7998	6.3689	1.8503	0.0766	44.6922	44.6156
Female	0.254179	−0.8202	−2.2300	0.5896	2.8196	0.4404	0.1075	1.8033	1.6957
Indian	0.035018	3.2425	0.2279	6.2570	6.0291	25.5970	1.2560	521.6705	520.4145
White	0.037299	2.6647	0.1568	5.1726	5.0159	14.3636	1.1697	176.3785	175.2087
Coloured	0.400166	1.3724	−1.8247	4.5695	6.3942	3.9447	0.1613	96.4963	96.3350
Science	0.000112	3.5484	1.7481	5.3487	3.6005	34.7578	5.7439	210.3271	204.5831

It is, however, important to interpret the OR correctly for models with multiple independent variables. Consider, for example, the *Indian* category of the variable *Race* in Combination 9. In the combination in which only the variable *Race* was included (Combination 2 in Table 8.43), the odds of an *Indian* student being successful at tertiary level over that of an *African* is 6.18, while for Combination 9 this ratio is 25.60. The reader may well question why this is the case. By including other variables in the multiple logistic regression model, as is the case with Combination 9 (see Table 8.50), one is again able to compare the category against the reference category, but now keeping the other independent (control) variables constant (that is, keeping them at the same level and thus neutralising their effect). To put this in perspective, of the eleven *Indian* students in the sample, nine studied Engineering, a study field which has already been shown to be associated with a far lower success rate. Thus, whereas in Combination 2 the *Study field* of the students was not considered, in Combination 9 the effect of *Study field* is neutralised. This explains why the odds of an *Indian* student being successful at tertiary level over that of an *African* student being successful at tertiary level is so much larger for Combination 9 than for Combination 2. The same kind of interpretation is applicable to all combinations which include multiple independent variables.

The final combination, Combination 10 (see Table 8.51), contains the three independent variables that show the most promise in terms of indicating those students most likely to be successful at tertiary level, namely *Tertiary institution*, *Race* and *Study field*. The results are near identical to those of Combination 9 with the respective categories that were previously statistically significant again being so. That is, the *University of Pretoria* is statistically different from the *University of the Witwatersrand*, *Indian* and *White* are statistically different from *African*, and *Science* is statistically different from *Engineering*, while controlling for each of the other variables, respectively.

TABLE 8.51: *The p-values, $\hat{\beta}$-coefficients, ORs, and their CIs for Combination 10 of Case Study C.*

			$\hat{\beta}$-CI:				OR CI:		
	p-value	$\hat{\beta}$	2.5%	97.5%	width	OR	2.5%	97.5%	width
Intercept	0.000020	−3.444	−5.025	−1.863	3.162	0.032	0.007	0.155	0.149
SU	0.585315	0.733	-1.899	3.365	5.264	2.081	0.150	28.921	28.771
UCT	0.859148	−0.230	−2.772	2.312	5.084	0.794	0.063	10.092	10.030
UP	0.027906	2.108	0.229	3.987	3.758	8.230	1.257	53.878	52.621
UJ	0.217444	1.537	−0.905	3.979	4.885	4.650	0.404	53.481	53.076
UKZN	0.865399	0.252	−2.667	3.172	5.838	1.287	0.069	23.847	23.777
Indian	0.012548	3.465	0.744	6.185	5.440	31.963	2.105	485.279	483.174
White	0.031167	2.541	0.230	4.853	4.623	12.697	1.259	128.100	126.842
Coloured	0.339744	1.341	−1.412	4.095	5.507	3.824	0.244	60.027	59.783
Science	0.000131	2.961	1.443	4.478	3.034	19.309	4.235	88.037	83.801

Random forest variable importance scores

The next step was to evaluate the variable combinations, listed in Table 8.36 and Table 8.39, based on their random forest variable importance scores. Combinations 5, 6, and 7 were not considered as they each only contain a single quantitative independent variable. The random forest variable importance score results for the remaining seven combinations are shown in Figures 8.49–8.57. Note that, unlike logistic regression, random forests are unaffected by multicollinearity. Therefore, the variable importance scores described below were determined from a model containing all categories, including the reference classes, for each combination.

The results for Combination 1 (see Figure 8.49) show that neither category of the variable *Gender* is given an importance score of more than zero. This is expected as up to now *Gender* has not shown any indication of being an important variable in terms of being able to predict tertiary success of prospective students.

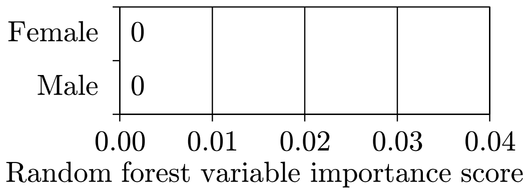


FIGURE 8.49: *Random forest variable importance scores for Combination 1 of Case Study C.*

Combination 2 (see Figure 8.50) contains the four categories of the *Race* variable. Those categories previously shown to be significant, *Indian* and *White*, along with the reference class of *African*, are given high random forest variable importance scores. The Category *African* is shown to be very important even when compared to the other important categories.

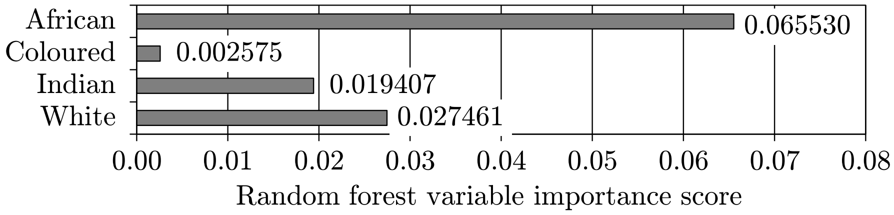


FIGURE 8.50: *Random forest variable importance scores for Combination 2 of Case Study C.*

For Combination 3 (see Figure 8.51) both the categories of the variable *Study field* are assigned very high random forest variable importance scores. This agrees with the previous analyses that also indicated *Study field* as being important with respect to predicting tertiary success.

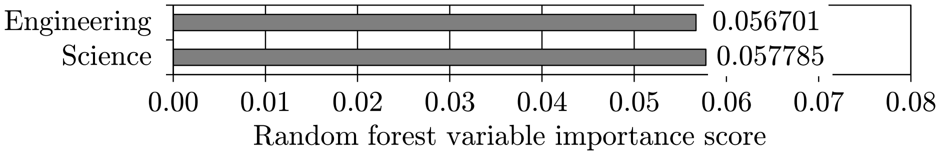


FIGURE 8.51: *Random forest variable importance scores for Combination 3 of Case Study C.*

For Combination 4 (see Figure 8.52), the variable *Tertiary institution* categories *University of Pretoria* and *University of the Witwatersrand* achieve high random forest variable importance scores. This result is to be expected since the *University of Pretoria* was the only category shown to be significant in the logistic regression analysis (the *University of the Witwatersrand* was the extreme category).

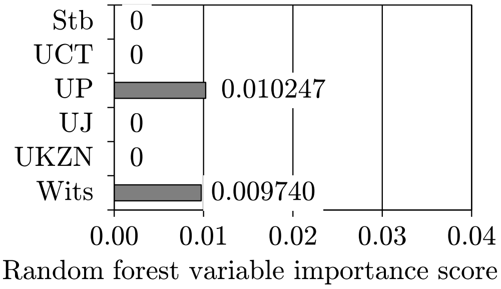


FIGURE 8.52: *Random forest variable importance scores for Combination 4 of Case Study C.*

Combination 8 (see Figure 8.53) contains the three high school variables *Grade 12 Mathematics*, *Grade 12 Physical science*, and *Grade 12 English.* It is slightly unexpected that *Grade 12 Mathematics* and *Grade 12 Physical science* achieve a reasonably high random forest variable importance score. This should, however, not come as a complete surprise as these two variables showed signs of perhaps being significant in the logistic regression analysis (see Table 8.49). Unlike for the $\hat{\beta}$-coefficients and ORs shown in the logistic regression results, it is not possible to identify whether or not an increase or decrease in the these school marks will result in a higher tertiary graduation success rate.

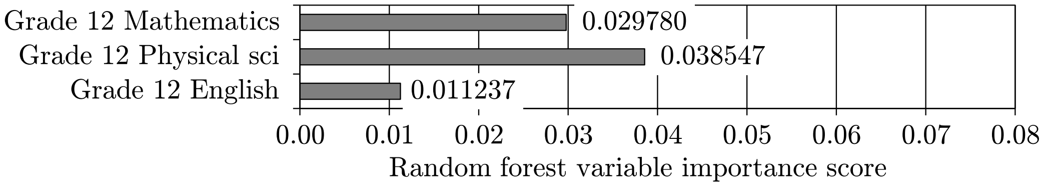


FIGURE 8.53: *Random forest variable importance scores for Combination 8 of Case Study C.*

For Combinations 9 and 10 two score plots are presented for each combination in Figures 8.54–8.55 (Combination 9) and 8.56–8.57 (Combination 10), respectively. The first plot for each of the two combinations is a model containing each qualitative variable as a single field or column, while in the second plots each category of the qualitative variables are partitioned into multiple dummy fields or columns.

The first plot for Combination 9 (see Figure 8.54), containing no dummy variables, identify *Race* and *Study field* as achieving the largest scores and *Tertiary institution*, *Grade 12 Mathematics*, and *Grade 12 Physical science* as also achieving reasonably high scores. These results correspond well with the random forest scores of the previous combinations that contained these five variables, where similar importance scores were assigned to each.

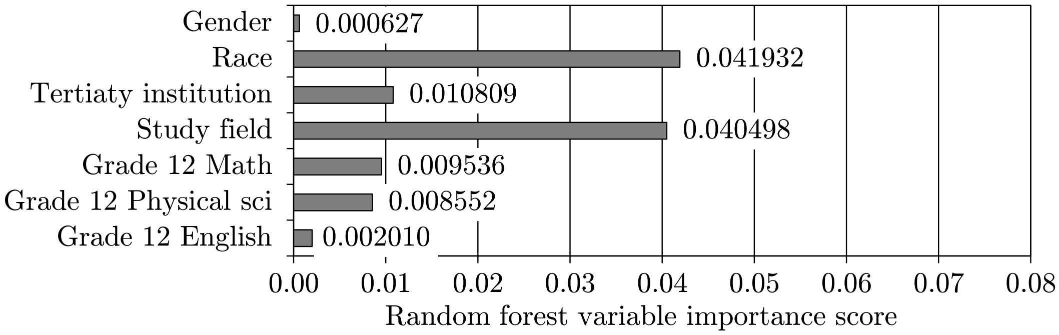


FIGURE 8.54: *Random forest variable importance scores for Combination 9 of Case Study C.*

The second plot for Combination 9 (see Figure 8.55), containing dummy variables, echoes the results of the previous plots that contained the same variables as in Combination 9.

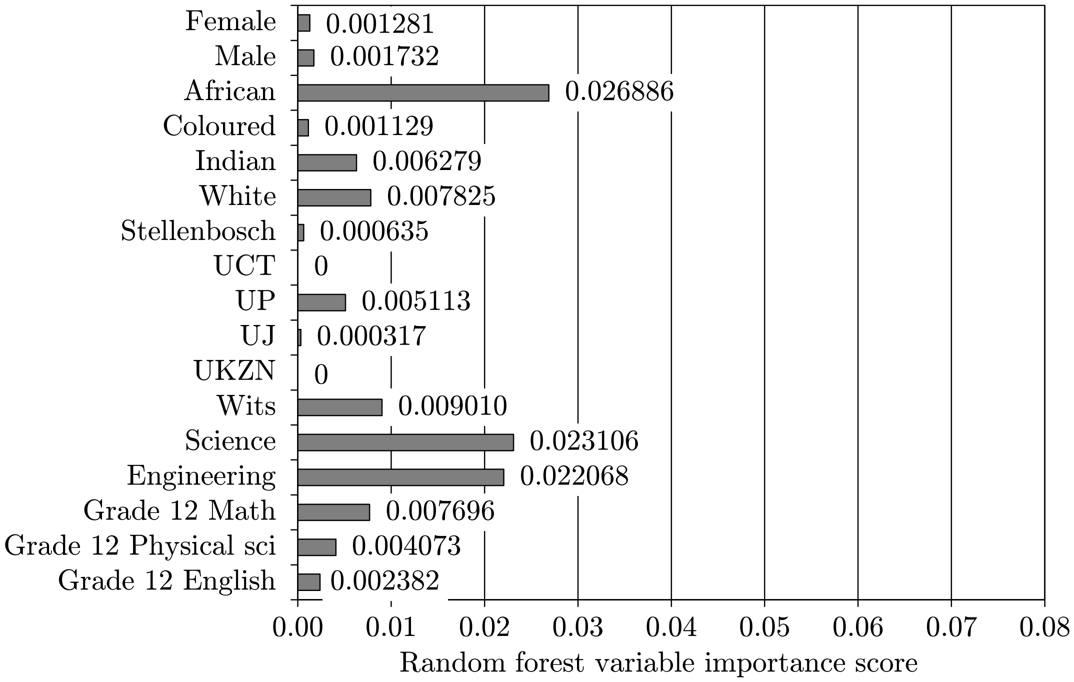


FIGURE 8.55: *Random forest variable importance scores for Combination 9 of Case Study C, including dummy variables.*

Again the categories of *African*, *Indian*, and *White* of the variable *Race* are identified as important with respect to predicting tertiary success. Similarly, the *Tertiary institution* categories *University of Pretoria* and the *University of the Witwatersrand* are identified as important with respect to being able to predict tertiary success. Also, both the categories *Science* and *Engineering* of the *Study field* variable are identified as important. Finally, *Grade 12 Mathematics* is shown as important predictors of tertiary success with *Grade 12 Physical science* less so, but still slightly important.

The two plots for the final combination, Combination 10 (Figures 8.56–8.57), exhibit random forest variable importance scores that echo the results of Combination 9. The only noticeable difference would appear to be that the *Coloured* category of *Race* has increased in importance score.

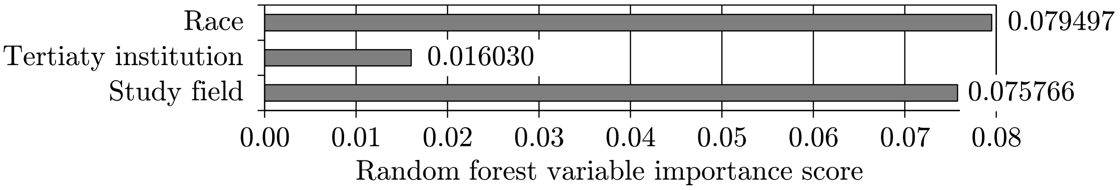


FIGURE 8.56: *Random forest variable importance scores for Combination 10 of Case Study C.*

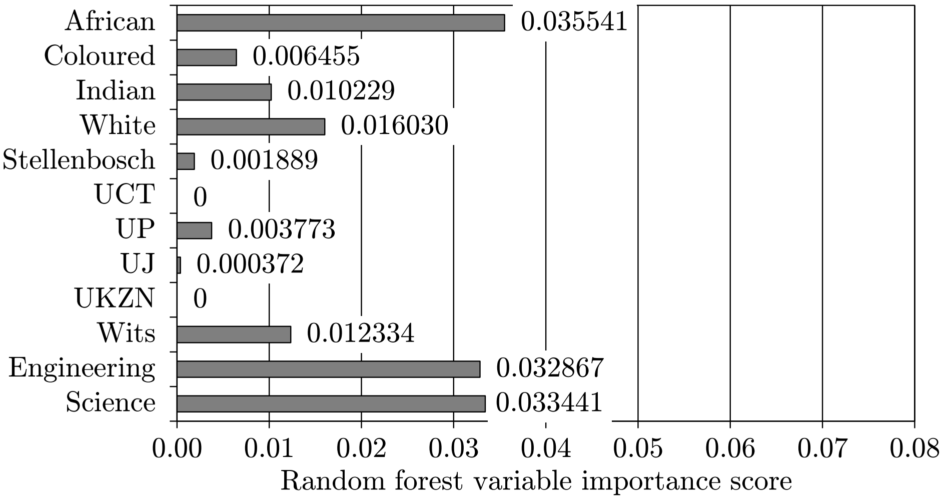


FIGURE 8.57: *Random forest variable importance scores for Combination 10 of Case Study C, including dummy variables.*

CART plots for the classification models

The third step was to evaluate the variable combinations of Tables 8.36 and 8.39 by means of CART plots. One plot was drawn up for each of the ten combinations, with the exception of Combination 1, as the root node did not partition into child nodes in this case. The nine resulting CART plots are shown in Figures 8.58–8.66. The same keys presented in Figure 8.3 may be used for the interpretation of nodes in the CART plots of this section.

The CART plot for Combination 2 (see Figure 8.58) produces expected results by partitioning the *Race* categories *African* and *Coloured* into a node with a high probability of failure, and the remaining categories *Indian* and *White* into a separate node with a reasonably high probability of tertiary success.

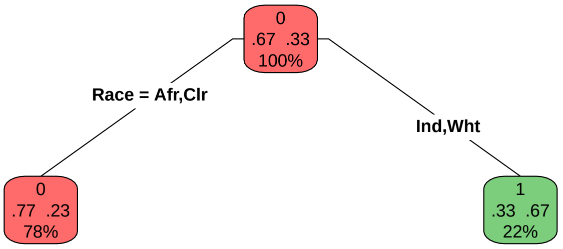


Figure 8.58: *CART plot for Combination 2 of Case Study C.*

For Combination 3 (see Figure 8.59), the CART plot contains three terminal nodes. The first (left-most) terminal node contains the *University of the Witwatersrand* with the lowest success proportion and the third (right-most) terminal node contains the *University of Pretoria* with the highest success proportion. Although the node containing only the *University of Pretoria* achieves a positive success proportion, the probability of obtaining tertiary success is still very close to 50% and so this category is not a good predictor of tertiary success.

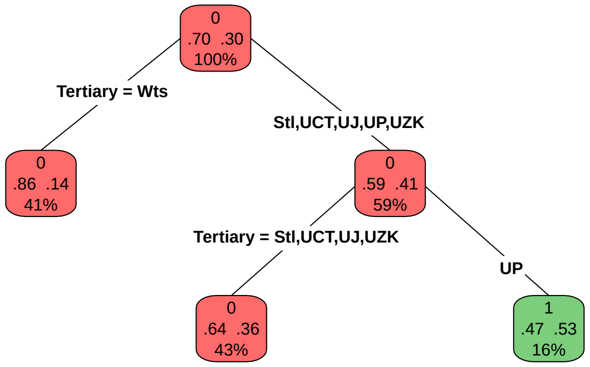


Figure 8.59: *CART plot for Combination 3 of Case Study C.*

The CART plot for Combination 4 (see Figure 8.60) also produces expected results by partitioning the *Study field* category *Engineering* into a node with a high probability of failure and the category *Science* into a separate node with a reasonably high probability of success. This single partition provides much insight into the success rates of the two study fields, with *Engineering* appearing to be associated with a much higher withdrawal rate.

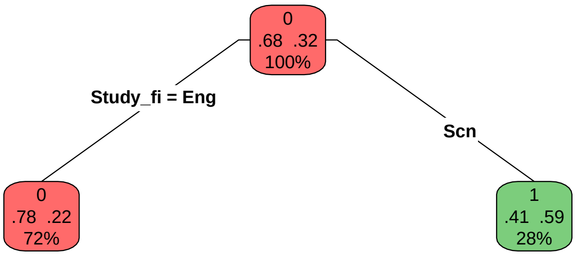


FIGURE 8.60: *CART plot for Combination 4 of Case Study C.*

The CART plot for Combination 5, shown in Figure 8.61, contains the single independent variable *Grade 12 Mathematics*. The results are peculiar in that it would seem that those students who achieved a *Grade 12 Mathematics* mark between 88% and 92% are unlikely to succeed at tertiary level, but those who obtained a *Grade 12 Mathematics* mark outside this bracket are far less likely to success at tertiary level. The above-mentioned observation would appear to indicate that no clear positive or negative trend exists between *Grade 12 Mathematics* and tertiary success.

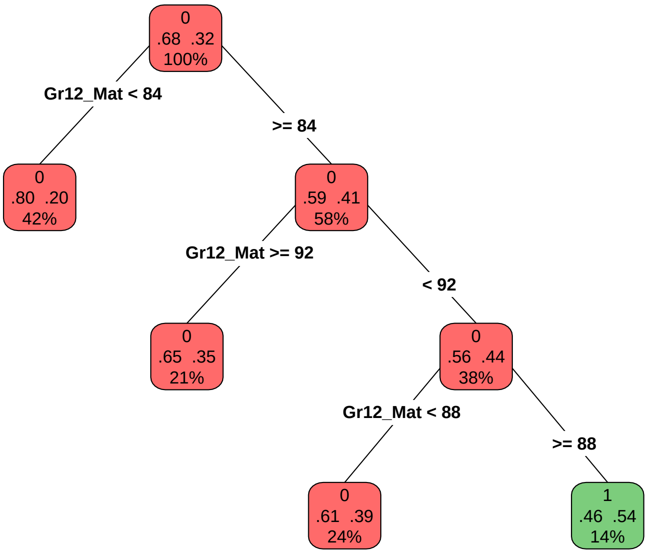


FIGURE 8.61: *CART plot for Combination 5 of Case Study C.*

As with *Grade 12 Mathematics*, the *Grade 12 Physical science* marks also do not seem to have a clear positive or negative relation to tertiary success. This is due to the CART plot (see Figure 8.62) indicating that within small, randomly placed, percentage margins the probability of success is higher than for other mark values.

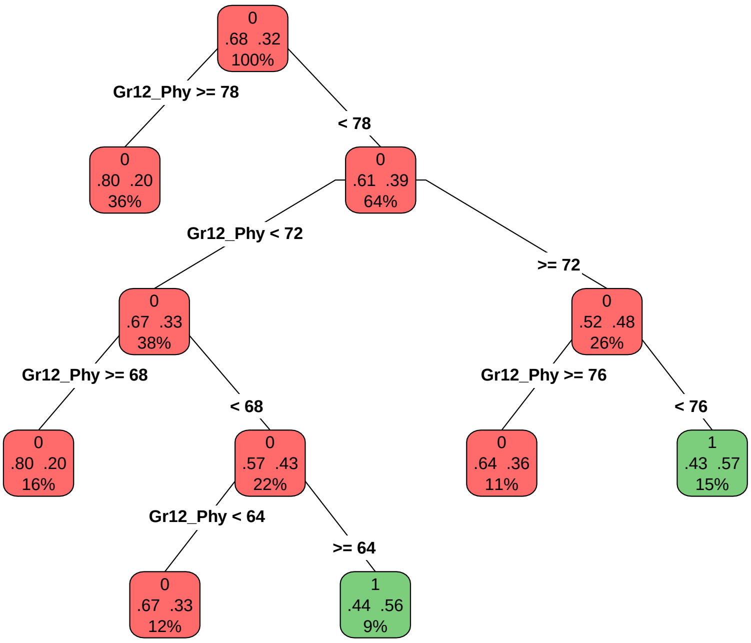


FIGURE 8.62: *CART plot for Combination 6 of Case Study C.*

The CART plot for Combination 7 (see Figure 8.63) contains only one partition. The probability of success of students within each of the first two child nodes are very similar to that of the root node, thus indicating that *Grade 12 English* is not a good predictor of tertiary success.

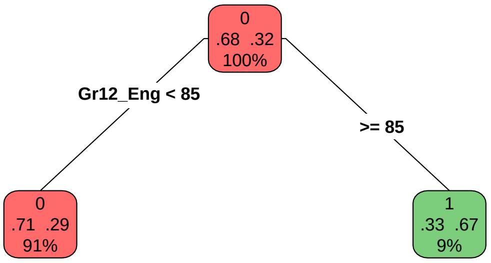


FIGURE 8.63: *CART plot for Combination 7 of Case Study C.*

Combination 8 (see Figure 8.64) produces interesting results. A low *Grade 12 English* mark together with a high *Grade 12 Physical science* mark seems to lead to a very low probability of success, while a low *Grade 12 English* mark, a low *Grade 12 Physical science* mark, and a high *Grade 12 Mathematics* mark seems to lead to a very high probability of success. Findings as specific as this are most likely only applicable to the specific sample considered, especially in the case of such a small sample, and should not be reported as general findings.

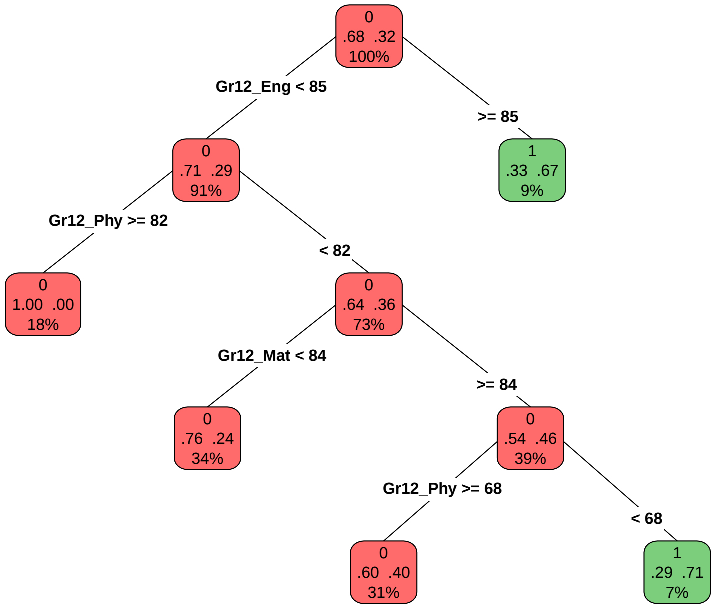


Figure 8.64: *CART plot for Combination 8 of Case Study C.*

Combination 9, the CART of plot which may be seen in Figure 8.65, contains all the available independent variables of Case Study C, listed in Table 8.35. Unsurprisingly the three variables that appear best-equipped to predict tertiary success, namely *Race*, *Study field*, and *Tertiary institution*, are the only variables that are partitioned in Combination 9. The interesting aspect to notice in this case, however, is the order in which the partitioning occurs. The CART considers *Race* the most influential variable in terms of predicting tertiary success, followed by *Study field*, and thirdly *Tertiary institution.* This is, however, expected as the random forest variable importance scores attributed to the variables of this combination (Figure 8.54) ranked the top three, in this order.

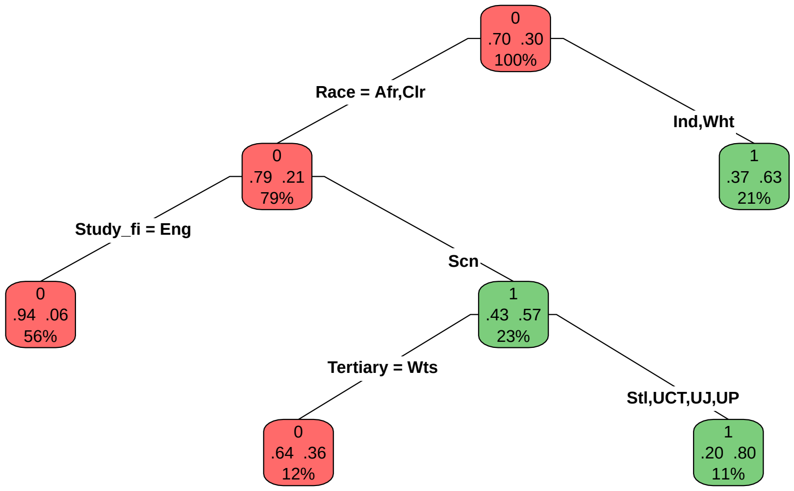


Figure 8.65: *CART plot for Combination 9 of Case Study C.*

The CART plot of Combination 10 (see Figure 8.66), is identical to that of Combination 9. This is not unexpected since Combination 10 excludes all of the variables that were not partitioned in the CART plot of Combination 9.

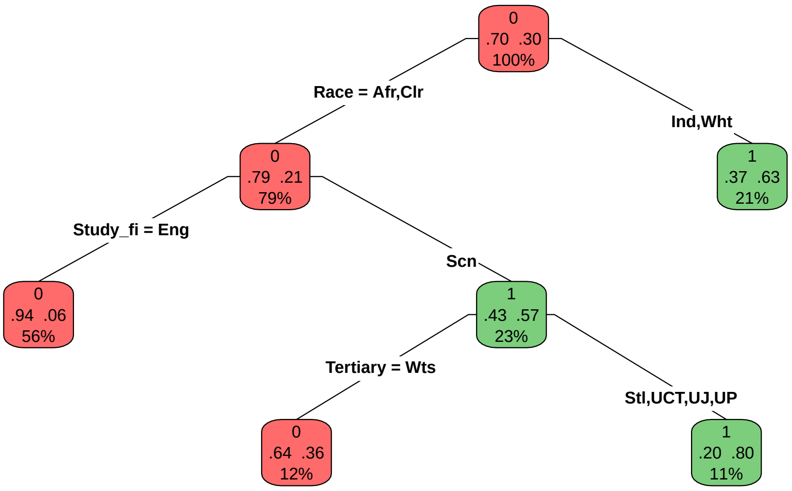


Figure 8.66: *CART plot for Combination 10 of Case Study C.*

Draw bar plots of variables of interest

The last, but crucial step, is to assess each individual independent variable by plotting the percentage of students within each category (for qualitative variables) and bin (for quantitative variables) who were successful in tertiary studies. These plots for the seven independent variables of Table 8.35 may be seen in Figures 8.68–8.73.

To emphasise why it is so vital to consider the above results in the context of these plots, consider again the ORs produced for Combination 3 containing the single independent variable of *Tertiary institution* shown in Table 8.44. It was shown that the odds of an individual who attended the *University of Pretoria* being successful over the odds of an individual who attended the *University of the Witwatersrand* being successful is 7.3. One may be tempted to incorrectly jump to the conclusion that the *Tertiary institution* category *University of Pretoria* will best indicate whether or not a student will be successful at his or her tertiary studies. The reason why this is not necessarily the case should become clear when considering Figure 8.67.

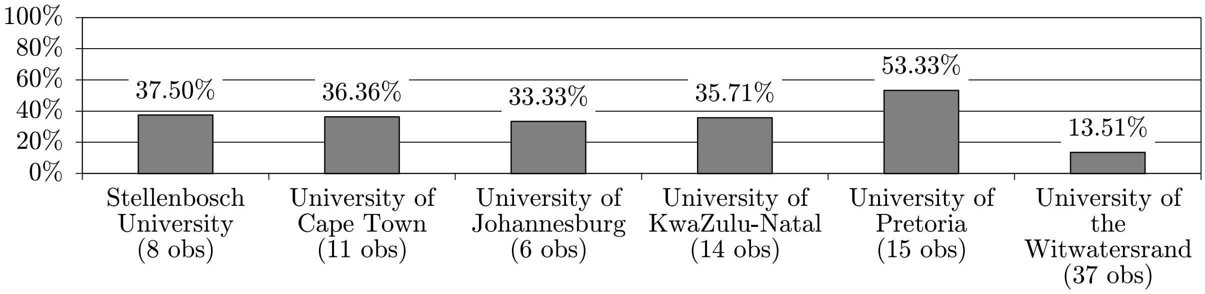


Figure 8.67: *Percentages of students who graduated successfully as a function tertiary institution in the sample of Case Study C.*

Students who attended the *University of Pretoria* were indeed more likely to succeed in their tertiary studies than students from the other tertiary institutions when compared against those who attended the *University of the Witwatersrand*. Notice, however, that in reality it is the most difficult to make accurate predictions about students who attended the *University of Pretoria* since their success rate is the closest to 50% at 53.33%. From the remaining *Tertiary institution* categories, the last category (the *University of the Witwatersrand*) seems to be an excellent predictor of tertiary success (or in this case failure).

By investigating Figure 8.68 it should be evident why the *Gender* field has hitherto been shown to have such low predictive ability, as its two categories (*Female* and *Male*) achieve near identical tertiary success rates.

The percentage of *Engineering* and *Science* students who graduated successfully, shown in Figure 8.69, validates what has been observed previously with respect to how the *Study field* variable is a good predictor. This is clearly due to the large difference between the success rates of the two categories.

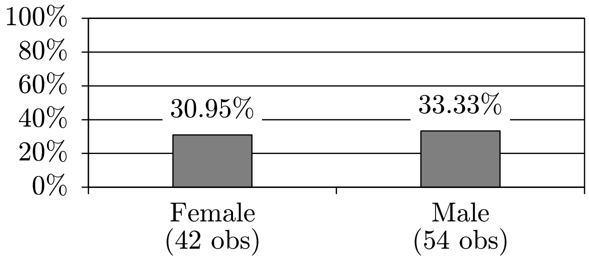


FIGURE 8.68: *Percentages of students who graduated successfully as a function gender in the sample of Case Study C.*

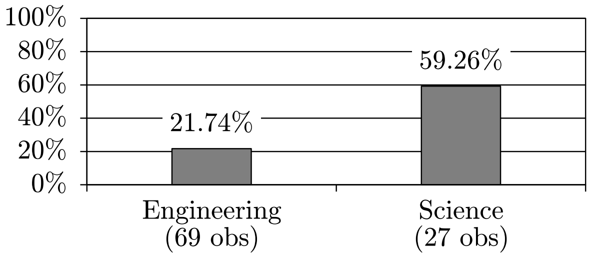


FIGURE 8.69: *Percentages of students who graduated successfully as a function study field in the sample of Case Study C.*

Displayed in Figure 8.70 is the tertiary success of students according to the four categories of the *Race* variable. This also confirms the previous findings that *Indian* and *White* students achieve the highest tertiary success rate and *African* and *Coloured* students the lowest. While reasonably confident results may be drawn about *African* students, less confident conclusions may be made about the other three races due to the low number of observations within each of those categories.

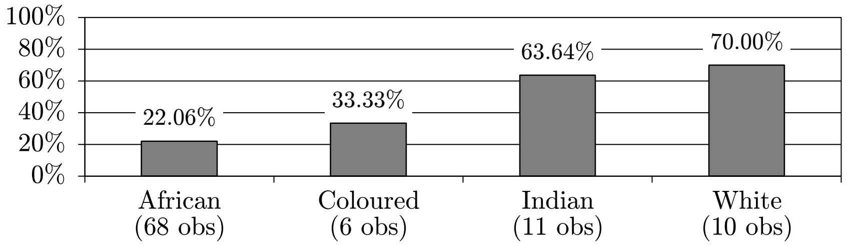


FIGURE 8.70: *Percentages of students who graduated successfully as a function race in the sample of Case Study C.*

Figure 8.71 contains the percentages of students who graduated as a function of *Grade 12 Mathematics*. Considering the last three bins, there appears to exist a slight positive trend between the students' *Grade 12 Mathematics* marks and tertiary success rate. This slightly contradicts the findings of the CART plot (see Figure 8.60) in which it was shown that a random segment of students, those who obtained between 88% and 92%, have the highest probability of success. This is likely due to the large size of the bins used for this bar plot, and the fact that those students who obtained a *Grade 12 Mathematics* mark between 88% and 92% are distributed over two bins.

The peculiar results of the CART plot of *Grade 12 Physical science* in Figure 8.62 pertaining to how there exists no clear positive or negative relation between the *Grade 12 Physical science* mark and the tertiary success of students, is again evident in Figure 8.72, considering the varying success rates of students over the bins.

The frequency bar plot of *Grade 12 English* is shown in Figure 8.73. Considering only the three central bins, as the two outer bins have very few observations, there appears to be no significant trend present between *Grade 12 English* marks and tertiary success. This also corroborates the results of the above analyses in which *Grade 12 English* was not identified as being an important variable in respect of predicting tertiary success.

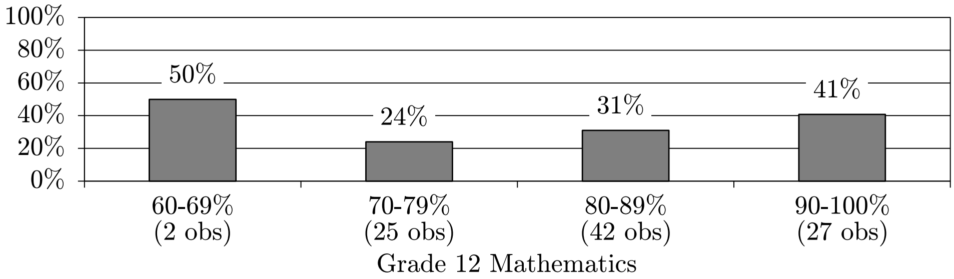


FIGURE 8.71: *Percentages of students who graduated successfully as a function Grade 12 Mathematics mark in the sample of Case Study C.*

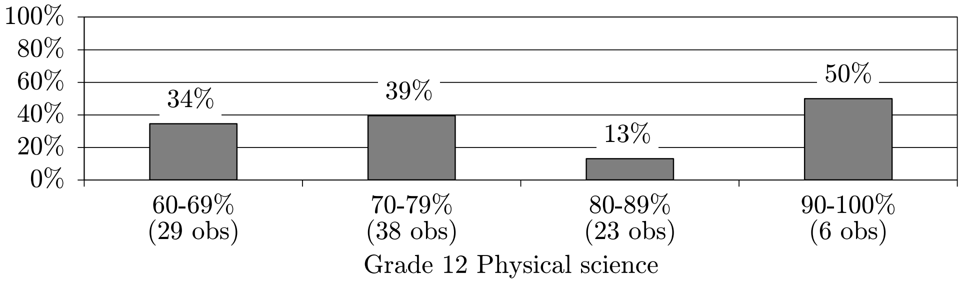


FIGURE 8.72: *Percentages of students who graduated successfully as a function Grade 12 Physical science mark in the sample of Case Study C.*

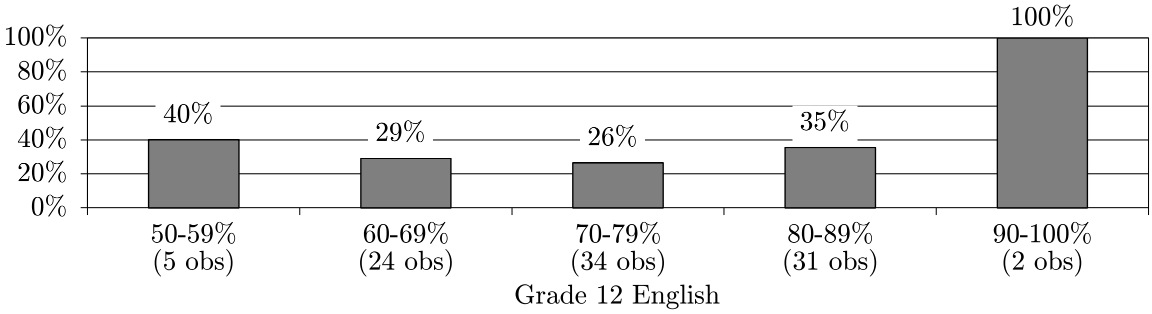


FIGURE 8.73: *Percentages of students who graduated successfully as a function Grade 12 English mark in the sample of Case Study C.*

8.4.8 Bootstrapped predictive accuracy assessment

The predictabilities of all ten combinations of Case Study C (their abilities to predict tertiary success) were evaluated by having each trained according to a bootstrapped sample, with replacement, and tested against the remaining OOB observations for 100 iterations. The average predictive performance of the weighted prediction produced by the DSS, for the OOB observations, for each of the twenty two combinations, may be seen in Table 8.52.

Upon initial inspection the results would appear very satisfactory due to the high predictive accuracies of the combinations, but the proportional sizes of the outcome classes of the sample data should be considered in conjunction with these results. The proportions of the observations of the sample that belong to each of the two outcome classes may be seen in the root node of any of the CART plots shown in Figures 8.58–8.66. For the sample used in all ten combinations, the outcome class with label “0” contains 68% of the observations. By taking this into consideration, the predictive performance of the combinations no longer appear as impressive. More specifically, Combinations 5, 6, and 7, which are concerned with the high school subjects *Grade 12 Mathematics*, *Grade 12 Physical science*, and *Grade 12 English*, are least predictive. Combination 1, containing the variable *Gender*, is also not very predictive. This result is unsurprising as these four combinations and variables have shown the least promise up to this point in the analysis in terms of possibly being able to predict tertiary success. The most predictive combination is Combination 10, containing the three variables *Race*, *Study field*, and *Tertiary institution*.

Table 8.52: *Average weighted prediction accuracy of selected combinations tested on 100 bootstrapped samples of the data of Case Study C.*

Combination number	Average weighted prediction accuracy	Combination number	Average weighted prediction accuracy
1	71.35%	6	70.25%
2	74.53%	7	70.83%
3	72.09%	8	69.85%
4	73.70%	9	75.57%
5	70.18%	10	77.17%

8.4.9 Report back of results

Two factors inhibited the interpretability and validity of the results of Case Study C. The sample considered was both small and skewed in the sense that a large portion of the sample had one of the two outcome labels. The skewness of the sample meant that often those categories and variables that were indicated as being significantly different from the reference classes during the logistic regression analyses did, in fact, only have a success rate of around 50%, and were thus not very predictive at all. The consequence of the small sample was large CI widths, indicating low precision. For these reasons, the results are only reported in very general terms (that is, no specific ORs are reported, reverting instead to a general overview of the important findings).

Despite the above restrictions, the following significant findings were identified. The three independent variables identified as clearly the most powerful at predicting tertiary success are *Race*, *Tertiary institution*, and *Study field*. The three high school marks of *Grade 12 Mathematics*, *Grade 12 Physical science*, and *Grade 12 English* performed disappointingly, especially in comparison to the above-mentioned three variables. These findings corroborate those of

Case Study B, namely that the difficulty of the study field and the standards of tertiary institutions for which students sign up has a far more significant effect on their probability of success. It would seem that this indicates the existence of extremely varying standards of qualifications between study fields and tertiary institutions within South Africa.

A major difference between the results of Case Study B and Case Study C is that the variable *Race* appeared much more able to successfully predict tertiary success of students in the latter. The reason for this may simply be that the two additional categories *Indian* and *White* were present in the sample of Case Study C. For both case studies it would seem that the success rates of *African* and *Coloured* students are not vastly different. These were the only categories present in the *Race* variable of Case Study B, and this may explain the low significance and variable importance scores. In Case Study C, where all four *Race* categories were present in the data, the difference in success rate of *African* and *Coloured* students *versus Indian* and *White* students resulted in *Race* being designated as a significant and good predictor of tertiary success. As in Case Study B, however, one cannot assume that the race of an individual will result in him or her being more successful at tertiary studies. One should rather keep in mind that there exist many other factors associated with race, such as *financial stability* and *family support structure*, that may have a large effect on a student's probability of tertiary success, especially within the South African context.

8.5 Case Study D

As explained in §7.7, the aim in this case study is to assess the capability and performance of the DSS using the same sample data as those considered in Case Studies B and C. The data of Case Study A are not used as it makes sense to rather use the larger sample from the same source, in this case that of Case Study B. The dependent variables to be considered and available independent variables have been declared in §7.7.

In order to consider the proposed DSS system a success, it should satisfy at least the following four criteria:

1. The system should achieve a better accuracy rate than the respective industry partners have achieved in the past using their current manual methods.

2. When considering the combined weighted prediction as a single base model competing with each of the other base models individually, it should on average achieve a higher accuracy than any of those five single base models the majority of the time.

3. At least some of the time the accuracy of the combined weighted prediction should exceed that of best base model's accuracy for particular runs or repetitions.

4. The ranking of the students produced by the MCDA method should make logical sense and agree with the predicted outcome of the students.

In order to assess the performance of the system according to the criteria above, it was necessary to analyse its performance in the context of multiple different scenarios. These scenarios were chosen as the variable combinations from Case Studies B and C that were found to be the most capable of predicting tertiary success. From Case Study B (see §8.3), Combinations 27, 28, 29, 30, 32, 33, and 34 were selected, while from Case Study C (see §8.4) Combinations 9 and 10 were chosen.

For each combination, 500 repetitions were run to assess the combinations' tertiary success predictive accuracy. Each repetition uses a different bootstrapped subset (with replacement) of the sample data to train the base models and test the trained models against the remaining OOB observations that were not selected to train the base models with. In addition to the randomness introduced by the bootstrapped re-sampling, a differently sized subset of sample data is used for the learning and validation each time. The size of the validation subset varies between 15% and 20% of the entire set of sample data available. The training subset comprises the remaining observations.

The results of the 500 repetitions for the seven combinations of case Study B and two combinations of Case Study C may be seen in Tables 8.53–8.54.

Table 8.53: *Performance of the accuracy of the weighted prediction for each of the ten combinations of the sample data of Case Study B.*

Combination number	Percentage of time the weighted prediction performed better than each of the base models:					Average weighted prediction accuracy	Percentage of time weighted prediction performed better than the best base model
	LR	CART	RF	C4.5	SVM		
27	69.55%	66.15%	57.37%	81.72%	80.47%	67.44%	16.80%
28	58.16%	84.73%	61.57%	71.21%	71.40%	67.02%	12.80%
29	70.10%	53.74%	72.39%	77.53%	75.70%	72.00%	8.60%
30	52.27%	90.30%	67.19%	78.66%	63.33%	69.49%	14.60%
32	75.54%	85.99%	89.36%	85.47%	69.11%	65.61%	10.70%
33	77.27%	51.68%	83.77%	73.07%	91.20%	68.77%	14.26%
34	54.63%	73.99%	61.14%	74.46%	86.32%	68.05%	9.20%
Average	65.36%	72.37%	70.40%	77.45%	76.79%	68.34%	12.42%

Table 8.54: *Performance of the accuracy of the weighted prediction for each of the two combinations of the sample data of Case Study C.*

Combination number	Percentage of time the weighted prediction performed better than each of the base models:					Average weighted prediction accuracy	Percentage of time weighted prediction performed better than the best base model
	LR	CART	RF	C4.5	SVM		
9	72.18%	93.94%	55.47%	92.78%	55.42%	76.21%	13.80%
10	69.44%	74.44%	50.00%	89.60%	94.58%	77.53%	0.20%
Average	70.81%	84.19%	52.74%	91.19%	75.00%	76.87%	7.00%

The tables provide three main pieces of information. The first (columns two to six) is the percentage of time that the weighted prediction performed better than each of the five base models. For example, when considering the first column and the first row of Table 8.53, one would be able to say that for the 500 repetitions of Combination 27, the weighted prediction had an accuracy better than that of the logistic regression model 69.55% of the time. Similarly, one may ascertain this statistic for the other four base models for each of the combinations. The second piece of information (provided in column seven) indicates the average accuracy of the weighted prediction over 500 repetition for each combination. The third and final piece of information (column eight) is the percentage of time the weighted prediction performed better than the best base model for a particular repetition.

The performance of the system is now compared against the four above-mentioned criteria in the following four subsections.

8.5.1 Improved accuracy compared to manual process

After the 500 applications of the DSS had been carried out, it was found that the average combined prediction accuracy of the system on the OOB validation sets was 68.34% for Case Study B and 76.87% for Case Study C.

Of the 1 101 students who remained in the sample of Case Study B after cleaning and who had received bursaries from the main industry partner over the past nine years, 643 have successfully graduated, while 458 were unsuccessful. This would imply a prediction accuracy of 58.40% for the main industry partner. Similarly, 32.29% of the students in the sample of Case Study C successfully graduated. To say that the predictive accuracy (with respect to tertiary success) of the two industry partners are 58.40% and 32.29%, respectively, would, however, be incorrect. The reason for this is that the majority of students in the two samples who started their tertiary studies in the past three to four years have all been unsuccessful at tertiary studies. This is because the sample only contains students who have either graduated or withdrawn, and many students who started their studies in the last three to four years will only be in the sample provided that they have withdrawn (unless they enrolled for a very short qualification). Students who started studying recently at tertiary level and are on track to graduate will not yet be in the sample. For this reason, the above calculation was redone, considering only those students who had started their tertiary studies in 2010 or earlier. The result is that the main industry partner (sample data of Case Study B) would seem to have an accuracy of 60.09% following their current manual process and similarly the secondary industry partner (sample data of Case Study C) achieved an accuracy of 51.79%.

As mentioned, tertiary success predictive accuracies of the weighted prediction for Case studies B and C is 68.34% and 76.87%, respectively. As both industry partners provide bursaries with the reasonable expectation that the students whom they support will graduate, it may be concluded that the first of the four above-mentioned criteria are satisfied as the proposed system produces better results than current manual attempt of both industry partners.

It should, however, be noted that the predictive accuracy of the weighted prediction was boosted by the skewed split of the two samples. The split in this case refers to the proportion of observations in the sample with the label of 1 *versus* those with the label 0. As shown above, the splits for the samples of Case Study B and Case Study C are 58.40% and 32.29% (or 67.71%), respectively. Despite this, the performance of the weighted prediction for Case Studies B and C are a 9.94% and a 9.16% improvement on these splits.

8.5.2 Improved accuracy compared to individual base models

The performance of the combined weighted prediction accuracy against the individual base models, averaged over the 500 runs for each combination, may be seen in columns two to six of Tables 8.53–8.54. The percentages displayed in the table were calculated by first excluding those cases for which the accuracy of the weighted prediction equalled that of a specific base model. Thus, the percentages only reflect occurrences for which the weighted prediction performed strictly better than each of the base models, individually.

From Tables 8.53–8.54 it is evident that the accuracy of the weighted predictions produced by the system outperforms each of the base models more than half the time, with the only exception being the method of random forests for Combination 10 of Case Study C, which was tied with the weighted prediction at 50%. It may thus be concluded that the weighted predictions of the system are better predictors than the predictions of any of the individual base models and so the second of the above four criteria is also satisfied.

8.5.3 Part time improved accuracy compared to best base model

The third criterion states that at least some of the time the accuracy of the weighted prediction should be greater than the best model's accuracy for specific runs. An expectation that this will always occur is unrealistic for the following reasons. Three outcomes are possible, namely either the accuracy of the weighted prediction is better than the best base model, equal to the best base model, or worse than the best base model. Very often a single base model performs very well compared to the rest and is assigned a large weighting. This results in the weighted prediction being above average, but still short of this best base model and is thus considered worse. Other times it is the case that all the base models produce the same accuracy, in which case the weighted prediction automatically achieves the same performance. The preferable case is that at least two base models produce a strong accuracy which allows the weighted prediction to achieve improved accuracy. One may expect that the addition of more good predictor models to the DSS would increase this percentage.

The last column of Tables 8.53–8.54 indicates the percentage of time that the accuracy of the weighted prediction exceeded that of the best base model. Although not high, it was explained why this is the case. To add more emphasis to why any improvement is significant, consider again the results of the example case considered in the walk-through in §6.5.5 and illustrated in §6.18–6.20. The validation set consisted of 145 observations which either yielded a predicted outcome of 1's by all the base models or 0's by all the base models for 112 of those observations. This implies that 77% of the alternatives had an obvious weighted prediction outcome and thus the weighted prediction only had the opportunity to really improve the prediction accuracy of the remaining 23% of the alternatives, in this case 33 alternatives. Of those 33 observations, 14 had only one of the five base models in disagreement, with the remaining four agreeing on the predicted outcome of an observation, resulting in the weighted prediction not being able to improve the prediction by combining the predictions for that specific alternative. This meant that the weighted prediction, in fact, only had the opportunity to improve the predictive accuracy for the remaining 13% of the observations. Obviously these percentages differ for different samples, but it should present the case that any predictive improvement noted by using the combined weighting, such as the 12.42% and 7.00% noted for Case Studies B and C, respectively, are that much more impressive. Based on the fact that 12.42% and 7.00% of the time the weighted prediction performed better than the best base model in Case Studies B and C, respectively, it may be concluded that the third criterion of the above four criteria is met.

After having trained the base models, an option that one may consider when predicting for new unseen alternatives is to only use predictions from the single best base model, based on the predictive ability of the base models identified during the training of the base models. In other words, to not combine the predictions of the multiple base models, but only use predictions from one base model, by assuming its solo prediction will perform better than the combined weighted prediction of all the base models for the new alternatives.

The danger is, of course, that by selecting only a single base model one effectively draws on only one 'opinion.' The power of the ensemble DSS is that one is able to draw on multiple, in the case of the concept demonstrator five, opinions. Although in theory one expects the predictive accuracy results on the OOB validation set to accurately reflect what would occur for a set containing new alternatives, it might not always be the case. In such cases, having multiple opinions will protect the user. That is, the use of strictly the best base model from the validation set might well produce the best results on a new data set, but the use of the weighted prediction which combines the base models is a far more stable and reliable alternative. To alter the DSS framework to operate like this will be easy and the final choice to do so will come down to personal preference. The DSS was purposefully not designed in this way for the reasons discussed above.

8.5.4 Sensible ranking

The fourth criterion requires that the ranking produced by the system should be logical and agree with the weighted predictions. In the system walk-through presented in §6.5.5, an in-depth discussion took place on how, for the majority of cases, the weighted prediction and the ranking agree that if moving in a descending order down the rank levels, the predictions change from 1's to 0's at some single point. This phenomenon makes logical sense since the higher-ranked students should be assigned a desired outcome and at some point down the ranking the students should be labelled as higher risks. Also pointed out, however, at times weighted predictions and rankings are produced that disagree on a few borderline observations, in which case the discrepancies should not be seen as an error, but rather as a valid second opinion and thus an asset. One should also consider the ranking valid based on its more than often exact relationship with the weighted prediction that obtained average accuracies of 68.34% and 76.87% for Case Studies B and C, respectively. The fourth criterion may therefore also be considered met.

8.6 Chapter summary

The four case studies that were introduced in Chapter 7 were performed in this chapter. The first of the four, Case Study A, was concerned with investigating the variable importance of three variables currently utilised by the main industry partner in its bursary application process. At the end of the case study, various recommendations were made to the industry partner based on the findings of the study.

In the following section, a nine-step roadmap was presented. The purpose of the roadmap was to provide a systematic approach which may be used to assess the variable importance and other trends of a sample, and this roadmap was followed in Case Studies B and C. One of the most important steps of the roadmap comprised using logistic regression models, random forest variable importance scores, CART plots, and frequency bar plots to investigate various statistical and machine learning models in the case studies. The analyses of Case Studies B and C followed thereafter.

Next, Case Study D was performed. This final case study made use of sample data from both Case Studies B and C to assess the capabilities of the DSS framework proposed in Chapter 5. It was found that the DSS functions effectively in the context of the case study data according to four reasonable performance criteria.

Part IV

Conclusion

CHAPTER 9

Thesis summary & appraisal

Contents

9.1 Thesis summary 245
9.2 Appraisal of thesis contributions 247

The purpose of this chapter is twofold: First, to present a summary of the work contained in this thesis (in §9.1) and secondly to offer an appraisal of the contributions of the thesis (in §9.2).

9.1 Thesis summary

The first part of this thesis, Part I: Literature study, comprises Chapters 2–4. The fundamentals of ISs and data preparation were presented in Chapter 2, in fulfilment of Objective I(i) of §1.5. In the opening section, §2.1, the various categories into which raw data may be classified were reviewed. Thereafter, the various steps of data preparation were presented in §2.2. In the next section, §2.3, multiple SDMs and tools were outlined which may be utilised to develop ISs, and specifically DSSs. The chapter closed in §2.4 with a review of the notions testing, validation, implementation, quality assurance and maintenance of an IS.

In Chapter 3, the fundamentals of statistical learning and machine learning were reviewed, in fulfilment of Objective I(ii). The chapter opened in §3.1 with a description of the core elements of statistical learning, including the relationship between input and output variables, regression *versus* classification, and data splitting methods such as CV and bootstrapping. In §3.2, an investigation followed of the possible reasons for the existence of correlation between variables. An argument was presented in §3.3 for why linear regression is not appropriate for classification. In the sections that followed, five different statistical and machine learning models that were applied later in this thesis were reviewed. They are logistic regression (§3.4), CART (§3.6), random forests (§3.7), the C4.5 algorithm (§3.8), and SVMs (§3.9). Additional sections and subsections were dedicated to a discussion on methods applicable to variable significance and importance analysis, namely logistic regression model evaluation (§3.5) focussed on the notion of log likelihood, the Wald statistic, the p-value, CIs, and OR; random forest variable importance scores (§3.7.1); and random forest outlier identification (§3.7.2). Various statistical and data assumptions underlying each of the five statistical learning and machine learning models of §3.5–3.9 were finally reviewed in §3.10.

In the final chapter of Part I, Chapter 4, the concepts of ensemble learning and MCDA were presented in fulfilment of Objective I(iii). The core concepts applicable to ensemble learning, as well as the mechanisms by which they may be formulated, were reviewed in §4.1. This was followed in §4.2 with a discussion on integration strategies for merging elements of an ensemble method. Particular emphasis was placed on the classification of these strategies as static integration strategies or dynamic integration strategies.

The second part of this thesis, Part II: Decision support system, comprises Chapters 5–6. In Chapter 5, a modular and generic DSS architecture capable of combining the predictions for alternatives produced by multiple statistical learning and machine learning models into a single prediction and a presentation of these alternatives within rank levels was proposed, in fulfilment of Objective III. An overview of the DSS framework was presented in §5.1, and this was followed by further explanations pertaining to specific components of the framework in §5.2–5.3.

The implementation of a concept demonstrator of the proposed DSS framework of Chapter 5 was outlined in Chapter 6, in fulfilment of Objective IV. The seven phases the SDLC were used as the foundation of this chapter. An outline was provided in §6.1 of how each of these seven phases was performed. Phases 1 (identifying problems, opportunities, and objectives) and 2 (determining human information requirements) were considered in §6.2, in fulfilment of Objective II(iii). Phase 3 (analysis of the system needs) was considered in §6.3, where the main industry partner's current bursary application process was interpreted and visually presented by means of a use case and activity diagram, in fulfilment of Objective IV(i). The design of the system, Phase 4 of the SDLC, was discussed in §6.4, encompassing the establishment of a system architecture, which in this thesis was presented in Chapter 5. Phase 5 involved building the system, and was documented in §6.5. This included the actual development of the proposed DSS architecture of Chapter 5 as a concept demonstrator in fulfilment of Objective IV(ii). Also included in Phase 5 was the selection of the specific components to include in the DSS architecture, as described in §6.5.2, in fulfilment of Objectives IV(iii)–(vi). The specific components selected were the five statistical and machine learning base models of logistic regression, CART, random forests, the C4.5 algorithm, and SVMs (along with mechanisms for assessing their statistical and data assumptions). The integration strategy adopted is an extension of weighted majority voting, and an MCDA outranking method known as ELECTRE III. The final part of Phase 5 entailed the creation of a user-friendly GUI so as to render the functions of the DSS concept demonstrator readily available to users, in fulfilment of Objective IV(vii).

The third part of this thesis, Part III: Case studies, comprises Chapters 7 and 8. In Chapter 7, the data obtained from two industry partners were presented and the aims of four case studies to be performed in the following chapter, based on these data, were discussed. The chapter opened in §7.1 with an elucidation of the data population that is of interest in this thesis, in relation to other tertiary students. A data sample was obtained from the main industry partner and presented in §7.2. These sample data were applicable to Case studies A and B of the following chapter. The aim and variable selection for these case studies were described in §7.3 and §7.4, respectively. Case study A was concerned with answering a specific question with respect to the importance of three score variables currently used in the bursary application process of the main industry partner, while Case Study B assessed all the variables of the sample so as to identify interesting trends and variables which may be used to best predict tertiary success. A third case study, Case Study C of the following chapter, had an identical aim (laid out in §7.6) to that of Case Study B, namely to investigate general trends and identify powerful predictive variables, but was based on sample data provided by the secondary industry partner, presented in §7.5. The performance of the DSS concept demonstrator of Chapter 6 was assessed in Case Study D, the aim of which was elaborated on in §7.7.

The purpose of Chapter 8 was to document four case studies, introduced in Chapter 7. The tertiary success predictive power of three score variables used by the main industry partner were investigated in Case Study A (§8.1). This was followed by a nine-step roadmap, presented in §8.2, outlining the steps which may be followed to analyse a data sample for interesting trends and to search for powerful predictive variables. The above roadmap was applied during the application of Case Studies B (§8.3) and C (§8.4). Case Study D was finally performed in §8.5 with a view to validate the DSS of Chapter 6 in the context of its ability to perform as required.

9.2 Appraisal of thesis contributions

The main contributions of this thesis are twofold: First, the identification of variables that are able to predict the tertiary study outcome of underprivileged students, and secondly the proposal of a predictive and ranking ensemble framework. These two contributions are assessed in more depth in this section.

9.2.1 Variable importance analysis and trends investigation

Three different case studies of Chapter 8 were dedicated to investigating trends and searching for variables that can best predict the tertiary study outcome of underprivileged students. These case studies were performed using sample data from two industry parters in §8.1 and §8.3–8.4.

The significance of the findings of these case studies is that the findings do not conform to traditional beliefs, such as that the Grade 12 Mathematics mark of an underprivileged student is a good indicator of his or her tertiary success. Instead, the findings contradict many of these traditionally held beliefs. More specifically, the results of Case Study A (§8.1), namely that the three score variables of the main industry partner are not valuable in respect of predicting tertiary success, was surprising to the main industry partner which had been using these scores for a number of years in its bursary application process. Similarly, the finding that an increase in high school subject marks and high school average marks leads to an increased withdrawal rate is counter-intuitive, and was not expected by the main industry partner. These findings were able to shed light on a very important trend of which the main industry partner was previously unaware. The main industry partner provided the following feedback based on the variable importance analysis and trend investigation of §8.1 and §8.3 [245]:

> *"The student's research has assisted our NGO greatly in making strategic decisions. The timing of the research happened to overlap with a major review of our programme activities by an external evaluator, and the research has, in many ways, confirmed our findings. It has also highlighted very interesting, and sometimes alarming trends, but ultimately this will aid us in adjusting our focus and our approach which will undoubtedly result in a higher success rate of our students. The student has produced an excellent piece of work which will certainly not remain a purely academic exercise. It has already been read and analysed by several decision makers and it will be used as a tool for advocacy, improvement and change."*

It is clear that the main industry partner was able to see the value in the findings reported in this thesis and will be able to appropriately adopt its bursary application process and criteria accordingly.

As mentioned in §1.4, numerous previous studies have been conducted, all aimed at investigating variable importance with respect to predicting academic success of students. As was emphasised, however, the results of those studies are rarely able to translate across different sample populations due to the vastly different backgrounds of the populations and the variables available for each of these studies. For these reasons, the findings reported in this thesis, related specifically to underprivileged South African tertiary students, are in fact unique.

9.2.2 Proposed decision support system framework

The proposed DSS ensemble framework, presented in Chapter 5, allows a user to select a dependent variable, multiple independent variables, various statistical learning and machine learning base models, and be presented with both an outcome prediction and a rank level for new alternatives, based on trends discussed in historical data.

The DSS may be considered very flexible from a user's perspective as a result of how it places him or her in control. The user is allowed to select a unique combination of inputs, as outlined above, and is also in control in the sense that he or she is provided with additional information, such as an assessment of the validity of various underlying models assumptions, but may override the system's recommendation. In this manner, the user may experiment with the system's inputs and observe the effects that these inputs have on the resulting output.

The DSS also embraces a modular type design, since a systems analyst may create an incarnation of the system by including and/or excluding new and/or different components to the ones presented in this thesis. The three categories of components he or she may manipulate are the statistical learning and machine learning base models, the weighting method used, and the MCDA outranking method employed within the system.

The performance of the proposed DSS was assessed in §8.5. It is not possible to measure its exact average performance accuracy, as the average accuracy of the DSS depends on the sample used and the variables selected. Despite this, it was demonstrated that while the DSS was effectively blind when predicting the outcome of the validation sets, the weighted prediction, if considered as a prediction model in itself, outperformed all of the five statistical learning and machine learning base models individually (on only one occasion for a single combination did it match the performance of a base model). In addition, the DSS presents a rank level for each alternative. These rank levels may used in conjunction with the weighted predictions and at times be considered a valid second opinion when disagreements on borderline cases are present between the weighted prediction and rank levels.

The proposed DSS was demonstrated to two industry partners, who are involved with the provision and management of bursaries to underprivileged individuals. The two partners provided feedback based on their impressions of its potential.

The first was Glenda Glover, a full-time employee and the past director of a company who now manages bursaries and other financial support of tertiary students at the main industry partner, and who provided the following comment [245]:

> *"I have found the work and presentations of the student enormously interesting and thought-provoking. The question he is trying to answer with his research — prediction of success for Higher Education students — is not only of importance to our company, but for the whole Higher Education sector and institutions funding students. In a context that broadens access and has limited resources such tools are, I believe, vital for decision making."*

In addition, an expert in the bursary provision industry with more than thirty years of experience in the South African context who is currently managing director of an NGO specialising in bursary and scholarship provision, Monique Adams, provided valuable feedback after viewing a short demonstration of the concept demonstrator of Chapter 6 [1]. Her initial thoughts were positive as she stated:

> "*I am very encouraged by your model and I hope that at some point in the future it can be used by organisations such as ourselves to make better decisions...*"

She has, however, also seen in her time in the industry that when results of the kind produced by the system are made available to decision makers, they have a tendency of jumping to conclusions and thus had the following very valid warning to issue with respect to the use of the system by stating that:

> "*My concern is that whilst I feel that your model can be extremely helpful it may be a dangerous tool in the hands of some who may be inclined to overuse it or use it to the exclusion of other methods ... I think whoever uses it will need to be carefully briefed and 'trained' to avoid misuse or over reliance.*"

It is a reality that anyone in a consultant role may often miss the real need of his or her client as well us not have a true understanding of the industry space for which they are providing a service. The above feedback was very valuable and should be seriously considered by any organisation who intends to introduce a DSS of this kind to their organisation so as to avoid the above-mentioned pitfalls.

In closing, it should be mentioned that, although the working of the DSS was demonstrated in the context of bursary allocation data, it may be also applied to other fields in which the prediction of alternatives is required. Other potential application areas of the DSS include the banking sector, in which the prediction of whether or not clients will default on their payments may be pursued, or the insurance sector, in which the prediction of whether or not a potential client is a high risk might be the aim.

CHAPTER 10

Further work

This thesis now closes with a discussion on a number of ideas for future work. Each section of this chapter contains possibilities for related future work based on the area considered in the section. In particular, suggestions for future work relate to the variable importance analysis and data trend investigations carried out in this thesis (in §7.4, §7.6, §8.3, and §8.4), the predictive component of the DSS proposed in this thesis (in §5.3), the integration strategy component of the DSS proposed in this thesis (also in §5.3), the ranking component of the DSS proposed in this thesis (also in §5.3), the performance assessments of the DSS proposed in this thesis (in §7.7 and §8.5), the graphical user interface implemented in this thesis (in Chapter 6), and the overall DSS framework proposed in this thesis (in Chapter 5).

10.1 The variable importance analyses and trend investigations

Three suggestions, related to the variable importance analyses and trend investigations performed, may be considered for future work.

Suggestion 1: *Obtain further applicable sample data.*

A consideration for the future is to obtain more sample data related to the performance of underprivileged tertiary students and perform similar variable importance analyses as those conducted in Case Studies A (§8.1), B (§8.3), and C (§8.4) in order to validate or invalidate the findings of Chapter 8.

Suggestion 2: *Investigate the relationship between additional high school subject marks and specific study fields.*

Other variable importance analyses, which may be performed if more applicable sample data are available, involves investigating the effect of more of the high school subject variables against specific study fields. Only a few of the high school subject variables were included in the combinations of Case Studies B and C due to a lack of available data.

Suggestion 3: *Investigate the variable importance of the same variables, but over different populations.*

Other topics which may be investigated include the difference in the predictive accuracy of the same variables as those considered in Chapters 7 and 8, but over two different populations. For example, sample data of middle class tertiary students may be selected as well, and the effect of specific high school subject marks between the two samples may be compared in respect of tertiary success. In such an analysis, the many confounding variables of difference in high school quality and family structure should, however, be considered.

10.2 The predictive component of the proposed DSS

Three suggestions related the to the predictive component of the proposed DSS, which may be pursued as future work, are presented in this section.

Suggestion 4: *Consider more and/or different base models in the DSS framework.*

An option for further consideration is to implement more and/or different base models in the DSS framework of Chapter 6. Examples of two such additional models are *discriminant analysis* and *artificial neural networks.*

Suggestion 5: *Further refine the setting of parameter values of the current base models.*

The setting of parameter values of the five statistical and machine learning base models included in this study may be further refined so as to improve their individual performance and thus the performance of the ensemble as a whole. More specifically, alterations, and possible improvements to these models may include experimenting with a different splitting criterion and/or pruning rules for CART and random forests. Another consideration may be to utilise a different kernel for the SVM.

Suggestion 6: *Only implement base models that can produce probabilities and do not convert these to binary outcomes.*

Another possible consideration for further investigation is to only include base models in the DSS that are able to predict probabilities, and then not to convert these probabilities to binary outcomes for the various base models. Combining such probabilities into a single expected probability of success for each student will, in itself, yield a ranking of alternatives, as opposed to only two groups produced by the current system.

10.3 The integration strategy component of the proposed DSS

The following two suggestions, related to the integration strategy component of the proposed DSS, may be investigated as future work.

Suggestion 7: *Investigate an alternative method of weighting in the current integration strategy.*

An extension of weighted majority voting was applied in this thesis as the integration strategy. Further work may be considered in the form of investigating a different type of extension of weighted majority voting, where the weights of each base model is adjusted in some unique manner, based on its predictive performance in respect of the validation set. It might even be possible to consider pursuing optimisation of this weighting for each sample over a number of iterations by determining which weighting produces the best final weighted prediction average of the alternatives in the validation set, but in this case care should be taken to avoid overfitting.

Suggestion 8: *Investigate an alternative integration strategy based on sensitivity or specificity results.*

A more ambitious consideration may be to attempt to employ a form of dynamic integration using the sensitivity and specificity predictive accuracy percentages achieved by each base model during the prediction for the validation set. It would be required that for each alternative of the new data set, a weight be assigned for each base model based on whether it predicts "1" or "0" for the alternative. The weighting may be based on either the base models' sensitivity or specificity performances in respect of the validation set, when "1" or "0" is predicted for the new alternative, respectively. That is, a different weighting may be assigned to a prediction based on whether it predicts that the student will succeed or withdraw, *i.e.* based on how good the base model performs in making positive (student will pass) or negative (student will fail) predictions. The final prediction for each of these new alternatives may then be combined *via* the distinct weights.

10.4 The ranking component of the proposed DSS

Two suggestions for future work related to the ranking component of the DSS proposed in this thesis are now suggested.

Suggestion 9: *Alter the parameter settings of the ELECTRE III method.*

The setting of ELECTRE III may also be considered for optimisation. This may well be most easily achieved by altering the ELECTRE III threshold values of q, p, and i, from the settings applied in this thesis discussed in §6.5.2.

Suggestion 10: *Consider the implementation of different and/or more ranking method(s).*

Another topic for investigation is to employ a different ranking method and/or more ranking method(s). In the case that more ranking methods are employed, a method for combining the rankings of the various methods should also be pursued. An example of alternative or additional ranking methods that may be considered in the context of the DSS of this thesis include PROMETHEE I and PROMETHEE II.

10.5 Performance assessments of proposed the DSS

The following two suggestions, concerned with the performance assessment of the proposed DSS, may also be considered for future work.

Suggestion 11: *Reassess the performance of the DSS implemented using alternative components or alternative component settings.*

A consideration for further investigation is to reassess the performance and capabilities of the proposed DSS once alternative components or alternative component settings have been included in the framework implemented in Chapter 6. That is, to reassess the performance a DSS that employs a different combination of statistical and machine learning base models, a different integration strategy, and a different ranking method, or applies different settings to the existing components.

Suggestion 12: *Reassess the predictive ability of the DSS using an alternative measure of accuracy.*

One may also pursue the topic of the DSS's performance further by employing a different measure of predictive accuracy, such as the *area under the curve* (AUC) value of the *receiver operating curve* (ROC), as opposed to the weighted average of the specificity and sensitivity percentages, as was done in this thesis.

10.6 The graphical user interface

Three suggestions for future work, related to the graphical user interface of the DSS proposed in Chapter 6, are presented in this section.

Suggestion 13: *Consider providing the user with more options through the addition supplementary controls to the GUI.*

A potentially very useful extension to the DSS concept demonstrator of Chapter 6 is to add more options and controls to the GUI, which allow for specific criteria and preferences to be selected by the user. Examples of such controls may be to allow the user to set a limit on the maximum percentage of students who may pursue tertiary studies in a single province. Other examples may include allowing the user to ascertain the effect of only selecting students who wish to pursue a qualification in a specific study field or who are of a specific gender. This may potentially be very informative as the user will be able to determine the effects on application selection as a results of changes applied to certain criteria.

Suggestion 14: *Investigate the possibility of adopting multiple objectives to suggest the 'best' alternatives.*

An element of optimising the selection of alternatives in respect of multiple objectives may also be investigated. Only the single objective, namely the maximisation of tertiary student success, was pursued in this study. Examples of further objectives may include consideration of the amount of funds available to an NGO over and above only attempting to maximise the number of tertiary successful students. This would require obtaining the costs of specific degrees at specific tertiary institutions. Such a multiple-objective problem will be interesting since the expected study period of each qualification, and quality of the qualifications, would also have to be considered in a trade-off fashion.

Suggestion 15: *Investigate software that will make a fully integrated implementation of a working DSS easier.*

A possible further consideration may include attempting to construct the DSS in such a manner that it can easily be used by bursary provision companies on their desktops, without requiring extensive or complex system integration. A possible approach toward realising such an aim would be to reconstruct the DSS in the `Python` programming language, with which the establishment of a desktop application is far easier than with `R`. The `Python` programming language also has a large number of statistical libraries available to programmers, much like `R`, but it is not as extensive.

10.7 The DSS framework

The following final suggestion for future work concerns the proposed DSS framework of Chapter 5.

Suggestion 16: *Consider including a component of variable importance and significance analysis in the DSS framework.*

A possible topic for further investigation in respect of the DSS framework may be to integrate a component of variable importance and significance analysis, much like what was done in Case Studies B (§8.3) and C (§8.4). Such an extension to the framework should, however, be approached with *extreme* caution, as incorrect conclusions may well be drawn from the statistical analyses results by an untrained user.

References

[1] ADAMS M, 2016, Managing Director at *Career Wise*, [Personal Communication], Contactable at `moniquea@careerwise.co.za`.

[2] AGRESTI A, 2007, *Introduction to categorical data analysis*, 2nd Edition, John Wiley & Sons, Hoboken (NJ).

[3] AGUINIS H, GOTTFREDSON RK & JOO H, 2013, *Best-practice recommendations for defining, identifying, and handling outliers*, Organizational Research Methods, **16(2)**, pp. 270–301.

[4] AHMED A, 2011, *Software project management: A process-driven approach*, 1st Edition, Taylor & Francis Group, Boca Raton (FL).

[5] AKBULUT Y, 2011, *Autonomous resource allocation in clouds: A comprehensive analysis of single synthesizing criterion and outranking based multiple criteria decision analysis methods*, PhD thesis, University of Victoria, Victoria.

[6] ALEXOPOULOS EC, 2010, *Introduction to multivariate regression analysis*, Hipokratia, **14(1)**, pp. 23–28.

[7] ALI S, HAIDER Z, MUNIR F, KHAN H & AHMED A, 2013, *Factors contributing to the student's academic performance — case study of Islamia University sub-campus*, American Journal of Educational Research, **1**, pp. 283–289.

[8] ALTMAN DG & BLAND JM, 1994, *Diagnostic tests 1: Sensitivity and specificity*, British Medical Journal, **308(6943)**, pp. 1552.

[9] AMOR SB, JABEUR K & MARTEL JM, 2007, *Multiple criteria aggregation procedure for mixed evaluations*, European Journal of Operational Research, **181(3)**, pp. 1506–1515.

[10] ANTONOGEORGOS G, PANAGIOTAKOS DB, PRIFTIS KN & TZONOU A, 2009, *Logistic regression and linear discriminant analyses in evaluating factors associated with asthma prevalence among 10- to 12-years-old children: Divergence and similarity of the two statistical methods*, International Journal of Pediatrics, **2009**, pp. 1–6.

[11] AUSTIN PC & TU JV, 2004, *Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality*, Journal of Clinical Epidemiology, **57(11)**, pp. 1138–1146.

[12] AVESON D & FITZGERALD G, 2006, *Methodologies for developing information systems: A historical perspective*, pp. 27–38 in AVISON D, ELLIOT S, KROGSTIE J & PRIES-HEJE J (EDS), *The past and future of information systems: 1976–2006 and beyond*, Springer, Boston (MA).

[13] AVISON D & FITZGERALD G, 2003, *Where now for development methodologies?*, Communications of the Association for Computing Machinery, **46(1)**, pp. 78–82.

[14] Ayyangar L, 2007, *Skewness, multicollinearity, heteroskedasticity — You name it*, Proceedings of the SAS Global Forum, Menlo Park (NJ), pp. 1–7.

[15] Baltzan P & Phillips A, 2012, *Business driven information systems*, 3rd Edition, McGraw-Hill, New York (NY).

[16] Baroni M & Evert S, 2010, *Descriptive statistics for continuous data*, (Unpublished) Technical Report, Center for Mind/Brain Science, University of Trento, Trento.

[17] Belton S & Stewart TS, 2002, *Multiple criteria decision analysis: An integrated approach*, 1st Edition, Kluwer Academic Publishers, Norwell (MA).

[18] Benyon D, 1987, *Towards a tool kit for the systems analyst*, The Computer Journal, **30(1)**, pp. 2–7.

[19] Berry W, 1993, *Understanding regression assumptions*, 1st Edition, SAGE Publications, Newbury Park (CA).

[20] Bewick V, Cheek L & Ball J, 2005, *Statistics review 14 — Logistics regression*, Critical Care, **9(1)**, pp. 112–118.

[21] Bhanot G, Alexe G, Venkataraghavan B & Levine AJ, 2006, *A robust metaclassification strategy for cancer detection from MS data*, PROTEOMICS, **6(2)**, pp. 592–604.

[22] Birch ER & Miller PW, 2007, *Tertiary entrance scores — Can we do better?*, Education Research and Perspectives, **34(2)**, pp. 1–23.

[23] Blakeslee S, 1990, *Lost on earth: Wealth of data found in space*, [Online], [Cited May 23rd, 2015], Available from `http://www.nytimes.com/1990/03/20/science/lost-on-earth-wealth-of-data-found-in-space.html`.

[24] Bollen K, Jackman R & Zhang H, 1990, *Modern methods of data analysis*, 1st Edition, SAGE Publications, Newbury Park (CA).

[25] Bouyssou D, 1990, *Building criteria: A prerequisite for MCDA*, pp. 58–80 in Bana e Costa CA (Ed), *Readings in multiple criteria decision aid*, Springer, Berlin.

[26] Bowers D, 2008, *Medical statistics from scratch: An introduction for health care professionals*, 2nd Edition, John Wiley & Sons, Chichester.

[27] Bramer M, 2007, *Principles of data mining*, 1st Edition, Springer, London.

[28] Brans JP & De Smet Y, 2016, *PROMETHEE Methods*, pp. 187–219 in Figueira JR, Ehrgott M & Greco S (Eds), *Multiple criteria decision analysis: State of the art surveys*, Springer, New York (NY).

[29] Breiman L, Friedman JH, Olshen RA & Stone CJ, 1984, *Classification and regression trees*, 1st Edition, Wadsworth, Belmont (CA).

[30] Breiman L, 2001, *Random forests*, Machine Learning, **45(1)**, pp. 5–32.

[31] Breiman L, 2001, *Statistical modelling: The two cultures*, Statistical Science, **16(3)**, pp. 199–215.

[32] Breiman L, 2003, *Two-eyed algorithms and problems*, Proceedings of the 14th European Conference on Machine Learning, Cavtat-Dubrovnik, p. 9.

[33] Brinkkemper S, 1996, *Method engineering: Engineering of information systems development methods and tools*, Information and Software Technology, **38(4)**, pp. 275–280.

[34] Burnstein I, 2002, *Practical software testing*, 1st Edition, Springer, New York (NY).

[35] Busato VV, Prins FJ, Elshout JJ & Hamaker C, 2000, *Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education*, Personality and Individual Differences, **29(6)**, pp. 1057–1068.

[36] Caputo B, Sim K, Furesjo F & Smola A, 2002, *Appearance-based object recognition using SVMs: Which kernel should I use?*, Proceedings of the NIPS workshop on statistical methods for computational experiments in visual processing and computer vision, Whistler.

[37] Castelli DM, Hillman CH, Buck SM & Erwin HE, 2007, *Physical fitness and academic achievement in third- and fifth-grade students*, Journal of Sport & Exercise Psychology, **29**, pp. 239–252.

[38] Çaydaş U & Ekici S, 2012, *Support vector machines models for surface roughness prediction in CNC turning of AISI 304 austenitic stainless steel*, Journal of Intelligent Manufacturing, **23(3)**, pp. 639–650.

[39] Ceriani L & Verme P, 2012, *The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini*, The Journal of Economic Inequality, **10(3)**, pp. 421–443.

[40] Chatterjee S & Hadi AS, 2006, *Regression analysis by example*, 5th Edition, John Wiley & Sons, Hoboken (NJ).

[41] Chen PP, 1976, *The entity-relationship model: Toward a unified view of data*, Association for Computing Machinery Transactions on Database Systems, **1(1)**, pp. 9–36.

[42] Collins M, Schapire RE & Singer Y, 2002, *Logistic regression, AdaBoost and Bregman distances*, Machine Learning, **48(1)**, pp. 253–285.

[43] Colson G & De Bruyn C, 1989, *Models and methods in multiple objective decision making*, Mathematical and Computer Modelling, **12(10-11)**, pp. 1201–1211.

[44] Cook R & Weisberg S, 1982, *Residuals and influence in regression*, 1st Edition, Chapman & Hall, New York (NY).

[45] Cortes C & Vapnik V, 1995, *Support vector networks*, Machine Learning, **20(3)**, pp. 273–297.

[46] Crammer K & Singer Y, 2002, *On the learnability and design of output codes for multiclass problems*, Machine Learning, **47(2)**, pp. 201–233.

[47] Curry LA, Nembhard IM & Bradley EH, 2009, *Qualitative and mixed methods provide unique contributions to outcomes research*, Circulation, **119**, pp. 1442–1452.

[48] Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J & Lawler JJ, 2008, *Random forests for classification in ecology*, Ecology, **88(11)**, pp. 2783–2792.

[49] Davis CE, Hyde JE, Bangdiwala SI & Nelson JJ, 1986, *An example of dependencies among variables in a conditional logistic regression*, pp. 140–147 in Moolgavkar SH & Prentice RL (Eds), *Modern statistical methods in chronic disease epidemiology*, John Wiley & Sons, New York (NY).

[50] De Montis A, De Toro P, Droste-Franke B, Omann I & Stagl S, 2016, *Criteria for quality assessment of MCDA methods*, pp. 99–133 in Getzener M, Spash C & Stagl S (Eds), *Alternatives for valuing nature*, Taylor & Francis Group, New York (NY).

[51] De Prel JB, Hommel G, Röhrig B & Blettner M, 2009, *Confidence interval or p-value?*, Deutsches Ärzteblatt International, **106(19)**, pp. 335–339.

[52] Debruyne M, 2009, *An outlier map for support vector machine classification*, The Annals of Applied Statistics, **3(4)**, pp. 1566–1580.

[53] Dennis A, Wixom BH & Tegarden D, 2005, *Systems analysis and design with UML version 2.0: An object-orientated approach*, 2nd Edition, John Wiley & Sons, Hoboken (NJ).

[54] Dietterich TG, 2000, *Ensemble methods in machine learning*, Proceedings of the Multiple classifier systems: First International Workshop, Cagliari, pp. 1–15.

[55] Dobson AJ, 1990, *An introduction to generalized linear models*, Chapman & Hall, London.

[56] Doucet J, 2009, *Components, purpose and function of information systems*, [Online], [Cited October 13th, 2015], Available from `http://goo.gl/w9nZJg`.

[57] Draper N & Smith H, 1996, *Applied regression analysis*, 1st Edition, John Wiley & Sons, New York (NY).

[58] Driscoll DL, Salib P & Rupert DJ, 2007, *Merging qualitative and quantitative data in mixed methods research: How to and why not*, Ecological and Environmental Anthropology, **3**, pp. 18–28.

[59] Duckworth AL & Seligman MEP, 2005, *Self-discipline outdoes IQ in predicting academic performance of adolescents*, Psychological Science, **16**, pp. 939–944.

[60] Duff A, Boyle E, Dunleavy K & Ferguson J, 2004, *The relationship between personality, approach to learning and academic performance*, Personality and Individual Differences, **36**, pp. 1907–1920.

[61] Ehrgott M, Figueira JR & Greco S, 2010, *Trends in multiple criteria decision analysis*, 1st Edition, Springer, Boston (MA).

[62] Eisenhauer T, 2012, *Free and open source software (FOSS) vs paid software for your social business needs*, [Online], [Cited December 10th, 2015], Available from `http://goo.gl/YyScDC`.

[63] Eldrandaly K, Hadi A & Abdelaziz AN, 2009, *An expert system for choosing the suitable MCDA method for solving a spatial decision problem*, Proceedings of the 9th lnternational Conference on Production Engineering, Design and Control, Alexandria, pp. 25–30.

[64] Esposito F, Malerba D & Semeraro G, 1997, *A comparative analysis of methods for pruning decision trees*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **19(5)**, pp. 476–491.

[65] Faul F, Erdfelder E, Buchner A & Lang A, 2009, *Statistical power analyses using GPower 3.1 — Tests for correlation and regression analyses*, Behavior Research Methods, **41(4)**, pp. 1149–1160.

[66] Figueira JR, Mousseau V & Roy B, 2016, *ELECTRE Methods*, pp. 155–185 in Figueira JR, Ehrgott M & Greco S (Eds), *Multiple criteria decision analysis: State of the art surveys*, Springer, New York (NY).

[67] Finlay PN & Wilson JM, 1987, *The paucity of model validation in operational research projects*, Journal of the Operational Research Society, **38(4)**, pp. 303–308.

[68] Fox J, 2008, *Applied regression analysis and generalized linear models*, 2nd Edition, SAGE Publications, Thousand Oaks (CA).

[69] French CD, 1995, *"One size fits all" database architectures do not work for DSS*, Proceedings of the 2nd ACM SIGMOD, San Jose (CA), pp. 449–450.

[70] Furnham A, Chamorro-Premuzic T & McDougall F, 2002, *Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance*, Learning and Individual Differences, **14**, pp. 49–66.

[71] Galathiya AS, Ganatra AP & Bhensdadia CK, 2012, *Classification with an improved decision tree algorithm*, International Journal of Computer Applications, **46(23)**, pp. 1–6.

[72] Garson GD, 2012, *Testing statistical assumptions*, (Unpublished) Technical Report, North Caroline State Univerisity, Statistical Associates Publishing, Asheboro (NC).

[73] Gasparini L & Lustig N, 2011, *The rise and fall of income inequality in Latin America*, pp. 691–714 in Ocampo JA & Ros J (Eds), *The Oxford handbook of Latin American economics*, Oxford University Press, Oxford.

[74] Gelinas U, Dull R & Wheeler P, 2010, *Accounting information systems*, 10th Edition, Cengage Learning, Stanford, (CA).

[75] Gelman A & Hill J, 2007, *Data analysis using regression and multilevel/hierarchical models*, 1st Edition, Cambridge University Press, New York (NY).

[76] Genuer R, Poggi JM & Tuleau-Malot C, 2010, *Variable selection using random forests*, Pattern Recognition Letters, **31(14)**, pp. 2225–2236.

[77] Gibilisco S, 2013, *Data store*, [Online], [Cited December 16th, 2015], Available from `http://goo.gl/48YAgP`.

[78] Gil-Aluja J, 2001, *Handbook of management under uncertainty*, Kluwer Academic Publishers, Dordrecht.

[79] Gislason PO, Benediktsson JA & Sveinsson JR, 2006, *Random Forests for land cover classification*, Pattern Recognition Letters, **27(4)**, pp. 294–300.

[80] Graham MH, 2003, *Confronting multicollinearity in ecological multiple regression*, Ecology, **84(11)**, pp. 2809–2815.

[81] Greenhalgh T, 1997, *How to read a paper: Statistics for the non-statistician*, British Medical Journal, **315(7104)**, pp. 422–425.

[82] Guitouni A & Martel JM, 1998, *Tentative guidelines to help choosing an appropriate MCDA method*, European Journal of Operational Research, **109(2)**, pp. 501–521.

[83] Gupta S, 2012, *The relevance of confidence interval and p-value in inferential statistics*, Indian Journal of Pharmacology, **44(1)**, pp. 143–144.

[84] Hand D, Mannila H & Smyth P, 2001, *Principles of data mining*, 1st Edition, MIT Press, Cambridge (MA).

[85] Hansen L & Salamon P, 1990, *Neural network ensembles*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **12(10)**, pp. 993–1001.

[86] Harrell FE, 2001, *Regression modelling strategies — With applications to linear models, logistic regression, and survival analysis*, 1st Edition, Springer, New York (NY).

[87] Hassan MR, Hossain MM, Bailey J, Macintyre G, Ho JWK & Ramamohanarao K, 2009, *A voting approach to identify a small number of highly predictive genes using multiple classifiers*, BMC bioinformatics, **10(1)**, pp. 1–12.

[88] Hastie T, Tibshirani R & Friedman J, 2009, *The elements of statistical learning*, 2nd Edition, Springer, New York (NY).

[89] Haykin S, 1998, *Neural networks: A comprehensive foundation*, 2nd Edition, Prentice Hall, Upper Saddle River (NJ).

[90] Hellerstein J, 2008, *Quantitative data cleaning for large databases*, (Unpublished) Technical Report, EECS Computer Science Division, University of California, San Fancisco (CA).

[91] Hilbe J, 2009, *Logistic regression models*, 1st Edition, Chapman & Hall, Boca Raton (FL).

[92] Hill T & Lewicki P, 2007, *Statistics — Methods and applications*, 1st Edition, StatSoft, Tusla (OK).

[93] Hogg RV & Tanis EA, 2005, *Probability and statistical inference*, 7th Edition, Pearson, Upper Saddle River (NJ).

[94] Hosmer D, Lemeshow S & Sturdivant R, 2013, *Applied logistic regression*, 3rd Edition, John Wiley & Sons, Hoboken (NJ).

[95] Hwang CL & Yoon K, 1981, *Multiple attribute decision making: Methods and applications*, Springer, Berlin.

[96] Immitzer M, Atzberger C & Koukal T, 2012, *Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data*, Remote Sensing, **4(9)**, pp. 2661–2693.

[97] Ishizaka A & Nemery P, 2013, *Multi-criteria decision analysis*, 1st Edition, John Wiley & Sons, Chichester.

[98] Ishwaran H, 2007, *Variable importance in binary regression trees and forests*, Electronic Journal of Statistics, **1**, pp. 519–537.

[99] James G, Witten D, Hastie T & Tibshirani R, 2013, *An introduction to statistical learning*, 1st Edition, Springer, New York (NY).

[100] Karami A, 2011, *Utilization and comparison of multi attribute decision making techniques to rank bayesian network options*, PhD thesis, University of Skövde, Skövde.

[101] Karatzoglou A, Meyer D & Hornik K, 2006, *Support vector machines in R*, Journal of Statistical Software, **15(9)**, pp. 1–6.

[102] Kedarisetti KD, Kurgan L & Dick S, 2006, *Classifier ensembles for protein structural class prediction with varying homology*, Biochemical and Biophysical Research Communications, **348(3)**, pp. 981–988.

[103] Kendall KE & Kendall JE, 2011, *Systems analysis and design*, 8th Edition, Pearson, Upper Saddle River (NJ).

[104] Khoonsari PE & Motie A, 2012, *A comparison of efficiency and robustness of ID3 and C4.5 algorithms using dynamic test and training data sets*, International Journal of Machine Learning and Computing, **2(5)**, pp. 540–543.

[105] Kim H & Loh WY, 2001, *Classification trees with unbiased multiway splits*, Journal of the American Statistical Association, **96**, pp. 598–604.

[106] Kimble C, 2010, *Information systems and strategy*, [Online], [Cited June 17th, 2015], Available from `http://goo.gl/dXQJLJ`.

[107] King JE, 2008, *Binary logistic regression*, pp. 358–384 in Osborn JW (Ed), *Best practices in quantitative methods*, SAGE Publications, Thousand Oaks (CA).

[108] Kleijnen J, 1995, *Verification and validation of simulation models*, European Journal of Operational Research, **82**, pp. 145–162.

[109] Klemens B, 2009, *Modelling with data: Tools and techniques for scientific computing*, Princeton University Press, Princeton (NJ).

[110] Knofczynski GT & Mundfrom D, 2007, *Sample sizes when using multiple linear regression for prediction*, Educational and Psychological Measurement, **68(3)**, pp. 431–442.

[111] Knupp J, 2014, *What is a web framework?*, [Online], [Cited December 6th, 2015], Available from `http://goo.gl/sTC2Kr`.

[112] Kornyshova E & Salinesi C, 2007, *MCDM techniques selection approaches: State of the art*, Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, Honolulu (HI), pp. 22–29.

[113] Kotsiantis SB, 2007, *Supervised machine learning: A review of classification techniques*, Informatica, **31**, pp. 249–268.

[114] Kovacic Z, 2010, *Early prediction of student success — Mining students' enrolment data*, Proceedings of the Informing Science & IT Education Conference, Wellington, pp. 647–665.

[115] Kutner M, Nachtsheim C & Neter J, *Applied linear regression models*, 3rd Edition, McGraw-Hill, New York (NY).

[116] Lane DM, Scott D, Hebl M, Guerra R, Osherson D & Zimmer H, 2016, *Introduction to statistics: An interactive ebook*, Rice University, Houston (TX).

[117] Law A & Kelton W, 1982, *Simulation modelling and analysis*, 1st Edition, McGraw-Hill, New York (NY).

[118] Lekdee K & Ingsrisawang L, 2010, *The empirical distribution of Wald, score, likelihood ratio, Hosmer-Lemershow (HL), and deviance for small sample logistic regression models*, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, pp. 1–5.

[119] Lewis RJ, 2000, *An introduction to classification and regression tree (CART) analysis*, Proceedings of the 310th Annual Meeting of the Society for Academic Emergency Medicine, San Fancisco, p. 14.

[120] Li C, Wang J, Wang L, Hu L & Gong P, 2014, *Comparison of classification algorithms and training sample sizes in urban land classification with landsat thematic mapper imagery*, Remote Sensing, **6(2)**, pp. 964–983.

[121] Liaw A & Wiener M, 2002, *Classification and Regression by randomForest*, R News, **2(12)**, pp. 18–22.

[122] Lin A & Fultz N, 2015, *Logistic regression with categorical predictors*, [Online], [Cited August 16th, 2014], Available from `http://goo.gl/U308Ck`.

[123] Lin HT, Lin CJ & Weng RC, 2007, *A note on Platt's probabilistic outputs for support vector machines*, Machine Learning, **68(3)**, pp. 267–276.

[124] Linkov I, Varghese A, Jamil S, Seager TP, Kiker G & Bridges T, 2005, *Multi-criteria decision analysis: A framework for structuring remedial decisions at contaminated sites*, pp. 15–54 in Linkov I & Ramadan AB (Eds), *Comparative risk assessment and environmental decision making*, Springer, Dordrecht.

[125] Little F & Silal S, 2014, *Risk factor response analysis*, (Unpublished) Technical Report, Departments of Public Health and Statistical Sciences, University of Cape Town, Cape Town.

[126] Louppe G, Wehenkel L, Sutera A & Geurts P, 2013, *Understanding variable importances in forests of randomized trees*, Proceedings of the Neural Information Processing Systems (NIPS) conference, Liege, pp. 431–439.

[127] Luna J, 2000, *Predicting student retention and academic success at New Mexico Tech*, PhD thesis, New Mexico Institute of Mining and Technology, New Mexico.

[128] Lund A & Lund M, 2013, *Laerd Statistics*, [Online], [Cited May 25th, 2015], Available from `https://statistics.laerd.com/aboutus.php`.

[129] Macan TH, Shahani C, Dipboye RL & Phillips AP, 1990, *College students' time management: Correlations with academic performance and stress*, Journal of Educational Psychology, **82(4)**, pp. 760–768.

[130] Majdi I, 2013, *Comparative evaluation of PROMETHEE and ELECTRE with application to sustainability assessment*, PhD thesis, Concordia University, Montreal.

[131] Malczewski J, 1999, *GIS and multi-criteria decision analysis*, 1st Edition, John Wiley & Sons, New York (NY).

[132] Malhotra MK, Sharma S & Nair SS, 1999, *Decision making using multiple models*, European Journal of Operational Research, **114(1)**, pp. 1–14.

[133] Mandel T, 1997, *The elements of user interface design*, 1st Edition, John Wiley & Sons, New York (NY).

[134] Manning CD, Raghavan P & Schütze H, 2009, *An introduction to information retrieval*, Cambridge University Press, Cambridge.

[135] Marascuilo L & McSweeney M, 1977, *Nonparametric and distribution-free methods for the social sciences*, Brooks Cole, Monterey (CA).

[136] Marschner IC, 2011, *glm2: Fitting generalized linear models with convergence problems*, The R Journal, **3/2(3)**, pp. 12–15.

[137] Marsh JL, Hutton JL & Binks K, 2002, *Removal of radiation dose response effects: An example of over-matching*, British Medical Journal, **325(7359)**, pp. 327–330.

[138] Marsland S, 2009, *Machine learning: An algorithmic perspective*, 1st Edition, Chapman & Hall, Danvers.

[139] Martel JM & Matarazzo B, 2005, *Other outranking approaches*, pp. 197–259 in Figueira JR, Ehrgott M & Greco S (Eds), *Multiple criteria decision analysis: State of the art surveys*, Springer, New York (NY).

[140] MathWorks, 2016, *Support vector machines for binary classification*, [Online], [Cited May 15th, 2010], Available from `http://goo.gl/0zmBim`.

[141] McDonald RA, Hand DJ & Eckley IA, 2003, *An empirical comparison of three boosting algorithms on real data sets with artificial class noise*, Proceedings of the 4th International Workshop: Multiple Classifier Systems, Guildford, pp. 35–44.

[142] Meloni JC, 2004, *Sams teach yourself PHP, MySQL and Apache all in one*, 2nd Edition, Sams Publishing, Carmel (CA).

[143] Menard S, 2002, *Applied logistic regression analysis*, 2nd Edition, SAGE Publications, Thousand Oaks (CA).

[144] Mendoza GA & Martins H, 2006, *Multi-criteria decision analysis in natural resource management: A critical review of methods and new modelling paradigms*, Forest Ecology and Management, **230(1-3)**, pp. 1–22.

[145] Merz CJ, 1996, *Dynamical selection of learning algorithms*, pp. 281–290 in Fisher D & Lenz HJ (Eds), *Learning from data: Artificial intelligence and statistics V*, Springer, New York (NY).

[146] Merz CJ, 1998, *Classification and regression by combining models*, PhD thesis, University of California, Irvine (CA).

[147] Milborrow S, 2011, *Plot rpart Models. An enhanced version of plot.rpart*, [Online], [Cited March 13th, 2016], Available from `http://CRAN.R-project.org/package=rpart.plot`.

[148] Miller, J, 2016, *Unequal scenes of Kya Sands*, [Online], [Cited April 17th, 2016], Available from `http://www.unequalscenes.com/`.

[149] Mingers J, 1989, *An empirical comparison of selection measures for decision-tree induction*, Machine Learning, **3(4)**, pp. 319–342.

[150] Modha J, Gwinnett A & Bruce M, 1990, *A review of information systems development methodology (ISDM) selection techniques*, Omega, **18(5)**, pp. 473–490.

[151] Mohammad RA, 2007, *Dilemma between the structured and object-orientated approaches to systems analysis and design*, Journal of Computer Information Systems, **46(3)**, pp. 32.

[152] Moisen GG, 2008, *Classification and regression trees*, Encyclopedia of Ecology, **(2008)**, pp. 582–588.

[153] Montclair State University, 2015, *Various kinds of data types known in statistics*, [Online], [Cited May 27th, 2015], Available from `http://pages.csam.montclair.edu/%7B~%7Dmcdougal/SCP/D%7B%5C_%7Dtypes.htm`.

[154] Montgomery D & Runger G, 2007, *Applied statistics and probability for engineers*, 4th Edition, John Wiley & Sons, Phoenix (AZ).

[155] Moore D, McCabe G & Bruce C, 2009, *Introduction to the practice of statistics*, 6th Edition, W H Freeman and Company, New York (NY).

[156] Mota P, 2013, *Comparative analysis of multi-criteria decision making methods*, PhD thesis, University of Almada, Almada.

[157] Mota P, Campos AR & Neves-Silva R, 2013, *First look at MCDM: Choosing a decision method*, Advances in Smart Systems Research, **3(2)**, pp. 25–30.

[158] Msilva V, 2016, *#FeesMustFall is just the start of change*, [Online], [Cited March 31st, 2016], Available from `http://mg.co.za/article/2016-01-20-fees-are-just-the-start-of-change`.

[159] Murphy KR & Myors B, 2014, *Statistical power analysis — A simple and general model for traditional and modern hypothesis*, 4th Edition, Taylor & Francis Group, New York (NY).

[160] Nisbet R, Elder J & Miner G, 2009, *Handbook of statistical analysis & data mining applications*, 1st Edition, Elsevier, London.

[161] Obbayi SR, 2011, *Types of database management systems*, [Online], [Cited December 12th, 2015], Available from `http://www.brighthub.com/internet/web-development/articles/110654.aspx`.

[162] O'Brien RM, 2007, *A caution regarding rules of thumb for variance inflation factors*, Quality & Quantity, **41(5)**, pp. 673–690.

[163] O'Connor MC & Paunonen SV, 2007, *Big five personality predictors of post-secondary academic performance*, Personality and Individual Differences, **43**, pp. 971–990.

[164] O'Dell J, 2010, *A beginner's guide to integrated development environments*, [Online], [Cited December 10th, 2015], Available from `http://mashable.com/2010/10/06/ide-guide/%7B%5C#%7DmMRYwRmN.uqh`.

[165] Opitz D & Maclin R, 1999, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research, **11**, pp. 169–198.

[166] Osborn JW & Waters E, 2002, *Four assumptions of multiple regression that researchers should always test*, Research & Evaluation, **8(2)**, pp. 1–5.

[167] Otto EP, 1979, *Success and failure in tertiary education, with reference to school attended — A re-examination*, Australian Journal of Teacher Education, **4(1)**, pp. 1–11.

[168] Pallant J, 2005, *SPSS survival manual*, 2nd Edition, McGraw-Hill, Sydney.

[169] Parikh R, Mathai A, Parikh S, Chandra Sekhar G & Thomas R, 2008, *Understanding and using sensitivity, specificity and predictive values*, Indian Journal of Ophthalmol, **56(1)**, pp. 45–50.

[170] Park HA, 2013, *An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain*, Journal of Korean Academy of Nursing, **43(2)**, pp. 154–164.

[171] Parke CS, 2013, *Essential first steps to data analysis: Scenario-based examples using SPSS*, SAGE Publications, Thousand Oaks (CA).

[172] Parker JDA, Summerfeldt LJ, Hogan MJ & Majeski SA, 2004, *Emotional intelligence and academic success: Examining the transition from high school to university*, Personality and Individual Differences, **36**, pp. 163–172.

[173] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M & Duchesnay E, 2011, *Scikit-learn: Machine learning in Python*, Journal of Machine Learning Research, **12**, pp. 2825–2830.

[174] Peduzzi P, Concato J, Kemper E, Holford T & Feinstein A, 1996, *A simulation study of the number of events per variable in logistic regression analysis*, Journal of Clinical Epidemiology, **49(12)**, pp. 1373–1379.

[175] Peng CJ, Lee KL & Ingersoll GM, 2002, *An introduction to logistic regression analysis and reporting*, Journal of Educational Research, **96(1)**, pp. 3–14.

[176] Pennsylvania State University, 2016, *Hypothesis tests & related Intervals*, [Online], [Cited March 8th, 2016], Available from `https://onlinecourses.science.psu.edu/stat504/node/39`.

[177] Pennsylvania State University, 2016, *Minimal cost-complexity pruning*, [Online], [Cited February 4th, 2016], Available from `https://onlinecourses.science.psu.edu/stat857/node/60`.

[178] Pintrich PR & De Groot EV, 1990, *Motivational and self-regulated learning components of classroom academic performance*, Journal of Educational Psychology, **82(1)**, pp. 33–40.

[179] Platt J, 2000, *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*, pp. 61–74 in Smola AJ, Bartlett PL, Schölkopf B & Schuurmans D (Eds), *Advances in large margin classifiers*, MIT press, Boston (MA).

[180] Poole C, 2001, *Low p-values or narrow confidence intervals: Which are more durable?*, Epidemiology, **12(3)**, pp. 291–294.

[181] Puuronen S, Terziyan VY & Tsymbal A, 1999, *A dynamic integration algorithm for an ensemble of classifiers*, Proceedings of the 11th International Symposium on Foundations of Intelligent Systems, London, pp. 592–600.

[182] Quinlan JR, 1992, *Learning with continuous classes*, Proceedings of the Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343–348.

[183] Quinlan JR, 1987, *Simplifying decision trees*, International Journal of Man-Machine Studies, **27(8)**, pp. 221–234.

[184] Quinlan JR, 1993, *C4.5: Programs for machine learning*, Morgan Kaufmann Publishers, San Francisco (CA).

[185] Rashidi S, Ranjitkar P & Hadas Y, 2014, *Modelling bus dwell time with decision tree-based methods*, Journal of the Transportation Research Board, **2418(1)**, pp. 74–83.

[186] Rice JA, 2007, *Mathematical statistics and data analysis*, 3rd Edition, Duxbury Press, Belmont (CA).

[187] Robnik-Šikonja M, 2004, *Improving random forests*, Proceedings of the 15th European Conference on Machine Learning, Pisa, pp. 359–370.

[188] Rogers M & Bruen M, 1998, *Choosing realistic values of indifference, preference and veto thresholds for use with environmental criteria within ELECTRE*, European Journal of Operational Research, **107(3)**, pp. 542–551.

[189] Rokach L & Maimon O, 2008, *Data mining with decision trees: Theory and applications*, World Scientific Publishing, River Edge (NJ).

[190] Roy B, 1985, *Méthodologie multicritère d'aide à la décision*, Economica, Paris.

[191] Roy B, 1991, *The outranking approach and the foundations of electre methods*, Theory and Decision, **31(1)**, pp. 49–73.

[192] Ruggieri S, 2002, *Efficient C4.5 (classification algorithm)*, IEEE Transactions on Knowledge and Data Engineering, **14(2)**, pp. 438–444.

[193] Rushton SP, Ormerod SJ & Kerby G, 2004, *New paradigms for modelling species distributions?*, Journal of Applied Ecology, **41(2)**, pp. 193–200.

[194] Saffari A, 2010, *Computer Vision and Pattern Recognition (CVPR)*, Proceedings of the 2010 IEEE Conference, San Francisco (CA), pp. 3570–3577.

[195] Sangeetha R & Kalpana B, 2011, *Performance evaluation of kernels in multiclass support vector machines*, International Journal of Soft Computing and Engineering, **1(5)**, pp. 138–145.

[196] SAS Institute Inc., 1999, *SAS OnlineDoc*, 8th Edition, SAS Institute Inc., Cary (NC).

[197] Satzinger JW, Jackson RB & Burd SD, 2007, *Systems analysis and design in a changing world*, 4th Edition, Thomson Course Technology, Boston, (MA).

[198] Sauro J, 2015, *Fundamentals of statistics: Nominal, ordinal, interval and ratio*, [Online], [Cited May 27th, 2015], Available from `http://www.usablestats.com/lessons/noir`.

[199] Shaliziz C, 2013, *Advanced data analysis from an elementary point of view*, 1st Edition, Cambridge University Press, Cambridge.

[200] Shelly G & Rosenblatt H, 2010, *Systems analysis and design*, 8th Edition, Gengage Learning, Boston (MA).

[201] Sheskin D, 2004, *Handbook of parametric and nonparametric statistical procedures*, 3rd Edition, CRC Press, Boca Raton (FL).

[202] Shi Y & Song L, 2015, *Spatial downscaling of monthly TRMM precipitation based on EVI and other geospatial variables over the tibetan plateau from 2001 to 2012*, Mountain Research and Development, **35(2)**, pp. 180–194.

[203] Shiotsu Y, 2014, *Web development 101: Top web development languages in 2014*, [Online], [Cited December 16th, 2015], Available from `https://www.upwork.com/blog/2014/03/web-development-101-top-web-development-languages-2014/`.

[204] Shmueli G, 2010, *To explain or to predict?*, Statistical Science, **25(3)**, pp. 289–310.

[205] Shofade OJS, 2011, *Considering hierarchical structure of criteria in ELECTRE decision aiding methods*, PhD thesis, Stanford University, Stanford (CA).

[206] Silva IS, 1999, *Cancer epidemiology — Principles and methods*, 1st Edition, International Agency for Research on Cancer, Lyon.

[207] Soureshjani MH & Kimiagari AM, 2013, *Calculating the best cut-off point using logistic regression and neural network on credit scoring problem - A case study of a commercial bank*, African Journal of Business Management, **7(16)**, pp. 1414–1421.

[208] Souza CR, 2010, *Kernel functions for machine learning applications*, [Online], [Cited June 11th, 2016], Available from `http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/`.

[209] Spicer J, 2005, *Logistic regression and discriminant analysis*, 1st Edition, SAGE Publications, Thousand Oaks (CA).

[210] Stärk KDC & Pfeiffer DU, 1999, *The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology — An example*, Intelligent Data Analysis, **3(1)**, pp. 23–35.

[211] Starkweather J, 2016, *Categorical variables in regression. Implementation and interpretation*, [Online], [Cited March 12th, 2014], Available from `http://it.unt.edu/research`.

[212] StatSoft Inc, 2016, *Popular decision tree: Classification and regression trees (CART)*, [Online], [Cited February 13th, 2016], Available from `http://www.statsoft.com/Textbook/Classification-and-Regression-Trees`.

[213] Steinberg L, Elmen JD & Mounts NS, 1989, *Authoritative parenting, psychosocial maturity, and academic success among adolescents*, Child Development, **60(6)**, pp. 1424–1436.

[214] Stephen DF, Welman JC & Jordaan W, 2004, *English language proficiency as an indicator of academic performance at a tertiary institution*, South African Journal of Human Resource Management, **2(3)**, pp. 42–53.

[215] Stodden V, 2006, *Model selection when the number of variables exceeds the number of observations*, PhD thesis, Universitat Rovira i Virgili, Stanford (CA).

[216] Stone D, Jarrett C, Woodroffe M & Minocha S, 2005, *User interface design and evaluation*, 1st Edition, Morgan Kaufmann Publishers, San Fancisco (CA).

[217] Strang G, 2009, *Introduction to linear algebra*, 4th Edition, Cambridge Press, Wellesley (MA).

[218] Strobl C, Boulesteix AL, Kneib T, Augustin T & Zeileis A, 2008, *Conditional variable importance for random forests*, BMC Bioinformatics, **9(307)**, pp. 1–11.

[219] Strobl C, Boulesteix AL, Zeileis A & Hothorn T, 2007, *Bias in random forest variable importance measures: Illustrations, sources and a solution*, BMC Bioinformatics, **8(25)**, pp. 1–21.

[220] Sullivan GM & Feinn R, 2012, *Using effect size — or why the p value is not enough*, Journal of Graduate Medical Education, **4(3)**, pp. 279–282.

[221] Szumilas M, 2010, *Explaining odds ratios*, Journal of the Canadian Academy of Child and Adolescent Psychiatry, **19(3)**, pp. 227–229.

[222] Tam CM, Tong TKL & Zhang H, 2007, *Decision making and operations research techniques for construction management*, 1st Edition, City University of Hong Kong Press, Hong Kong.

[223] Teghem J, Delhaye C & Kunsch PL, 1989, *An interactive decision support system (IDSS) for multicriteria decision aid*, Mathematical and Computer Modelling, **12(10-11)**, pp. 1311–1320.

[224] The World Bank, 2014, *2014 Gini index world map, income inequality distribution by country*, [Online], [Cited February 10th, 2015], Available from `https://commons.wikimedia.org/wiki/File:2014_Gini_Index_World_Map,_income_inequality_distribution_by_country_per_World_Bank.svg`.

[225] Thimbleby H, Blandford A, Cairns P, Curzon P & Jones M, 2002, *User interface design as systems design*, pp. 281–301 in Faulkner X, Finlay J & Détienne F (Eds), *People and Computers XVI - Memorable Yet Invisible: Proceedings of Human Computer Interaction 2002*, Springer, London.

[226] Timofeev R, 2004, *Classification and regression trees (CART) — Theory and applications*, PhD thesis, Humboldt University, Berlin.

[227] Tonidandel S, Lebreton JM & Johnson JW, 2009, *Determining the statistical significance of relative weights*, Psychological methods, **14(4)**, pp. 387–399.

[228] Trochim W, 2006, *Measurement error*, [Online], [Cited October 20th, 2015], Available from `http://www.socialresearchmethods.net/kb/measerr.php`.

[229] Trochim W, 2006, *Types of relationships*, [Online], [Cited October 21st, 2015], Available from `http://www.socialresearchmethods.net/kb/relation.php`.

[230] Tsymbal A, Pechenizkiy M, Cunningham P & Puuronen S, 2008, *Dynamic integration of classifiers for handling concept drift*, Information Fusion, **9(1)**, pp. 56–68.

[231] Tsymbal A, Pechenizkiy M, Puuronen S & Patterson DW, 2003, *Dynamic integration of classifiers in the space of principal components*, Proceedings of the 7th East European Conference: Advances in databases and information systems, Dresden, pp. 278–292.

[232] Tu YK, Kellett M, Clerehugh V & Gilthorpe MS, 2005, *Problems of correlations between explanatory variables in multiple regression analyses in the dental literature*, British Dental Journal, **199(7)**, pp. 457–461.

[233] Tzeng GH & Huang JJ, 2011, *Multiple attribute decision making: Methods and applications*, Taylor & Francis Group, Boca Raton (FL).

[234] UNESCO, 2015, *Introducing UNESCO*, [Online], [Cited February 10th, 2015], Available from `http://en.unesco.org/about-us/introducing-unesco`.

[235] Urhuogo I, Vann V & Chandan H, 2012, *Information systems maintenance: The application of total quality management construct*, Journal of Business Studies Quarterly, **3(3)**, pp. 1–15.

[236] Van der Ploeg T, Austin PC & Steyerberg EW, 2014, *Modern modelling techniques are data hungry — A simulation study for predicting dichotomous endpoints*, BMC Medical Research Methodology, **14(1)**, pp. 137–149.

[237] Van Heerden B, Aldrich C & Du Plessis A, 2008, *Predicting student performance using artificial neural network analysis*, Medical Education, **42(5)**, pp. 516–517.

[238] Venables WN & Ripley BD, 2002, *Modern Applied Statistics with S*, Springer, New York (NY).

[239] Vicki SL, 2015, *Introduction to agile methodologies*, [Online], [Cited March 26th, 2014], Available from `http://www.umsl.edu/%7B~%7Dsauterv/analysis/6840%7B%5C_%7Df09%7B%5C_%7Dpapers/Nat/Agile.html`.

[240] Wackerly DD, Mendenhall W & Scheaffer RL, 2008, *Mathematical statistics*, 7th Edition, Brooks Cole, Belmont (CA).

[241] Walters SA, Broady JE & Hartley RJ, 1994, *A review of information systems development methodologies*, Library Management, **15(6)**, pp. 5–19.

[242] Watt A & Eng N, 2014, *Database design*, 2nd Edition, Pressbooks, Montreal.

[243] Weisstein EW, 2016, *Hypothesis testing*, [Online], [Cited April 28th, 2016], Available from `http://mathworld.wolfram.com/HypothesisTesting.html`.

[244] Welling M, 2010, *Support vector machines*, (Unpublished) Technical Report, University of Toronto, Toronto.

[245] Wesson A, 2016, Communications Manager at *Rural Education Access Programme*, [Personal Communication], Contactable at `anel@reap.org.za`.

[246] Winston W, 2004, *Operations research: Applications and algorithms*, 4th Edition, Brooks Cole, Belmont (CA).

[247] Woods K, Kegelmeyer Jr. WP & Bowyer K, 1997, *Combination of multiple classifiers using local accuracy estimates*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **19(4)**, pp. 405–410.

[248] Xu L, Krzyzak A & Suen CY, 1992, *Methods of combining multiple classifiers and their applications to handwriting recognition*, IEEE Transactions on Systems, Man, and Cybernetics, **22(3)**, pp. 418–435.

[249] Yang P, Ho JWK, Zomaya AY & Zhou BB, 2010, *A genetic ensemble approach for gene-gene interaction identification*, BMC bioinformatics, **11(1)**, pp. 524–538.

[250] Yang P, Hwa Yang Y, B Zhou B & Y Zomaya A, 2010, *A review of ensemble methods in bioinformatics*, Current Bioinformatics, **5(4)**, pp. 296–308.

[251] Yang P, Zhou BB, Zhang Z & Zomaya AY, 2010, *A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data*, BMC bioinformatics, **11(1)**, pp. 1–12.

[252] Yohannes Y & Hoddinott J, 1999, *Classification and regression trees: An Introduction*, (Unpublished) Technical Report, International Food Policy Research Institute, Washington (DC).

[253] Zhu X, 2010, *Advanced natural language processing: Support vector machines*, (Unpublished) Technical Report, University of Wisconsin-Madison, Madison (WI).

Part V

Appendices

APPENDIX A

Communication registry

Table A.1 contains a registry of communication that took place between the author and various industry partners, but predominantly the main industry partner. From the registry it should be clear that the communication with the main industry partner was continuous.

Table A.1: *Registry of communication with main industry partners.*

Date	Location	Medium used	Purpose
12-Nov-14	N.A.	E-mail	Main industry partner provided basic company information and requirements in preparation for the first face-to-face meeting.
19-Nov-14	Main industry partner's head office	Face-to-face meeting	First meeting with main industry partner contact, director, database manager, and other important members to establish the feasibility of the project, to determine their requirements and propose a preliminary solution falling within the capabilities of the author to solve.
12-Feb-15	Main industry partner's head office	Face-to-face meeting	Second face-to-face meeting with main industry partner to further discuss and refine the project scope.
18-Feb-15	N.A.	E-mail	Provided main industry partner with project proposal created to meet their needs based on the discussion held on 12 February 2015.
24-Feb-15	N.A.	E-mail	Received feedback on the project proposal, which is approved.
26-Feb-15	N.A.	E-mail	Finalised *non-disclosure agreement* (NDA) and forward it to main industry partner.
18-Mar-15	N.A.	E-mail	NDA signed by main industry partner.
20-Mar-15	N.A.	E-mail	Screen shots received of data that can be provided by the main industry partner.
23-Mar-15	N.A.	E-mail	From the screen shots, data fields were selected to populate an `MS Excel` file. The names of these fields were sent to the main industry partner.
24-Mar-15	N.A.	E-mail	First sample data received (in an `MS Excel` spreadsheet) from main industry partner.

26-Mar-15	N.A.	E-mail	Receive annual statistics and report of main industry partner provided for further information about their NGO.
20-Apr-15 – 23-Apr-15	N.A.	E-mail	Main industry partner provided clarity about uncertainty pertaining to the data.
29-Apr-15	N.A.	E-mail	Date and time organised for meeting with database manager.
04-May-15	N.A.	E-mail	Second data sample provided by main industry partner, containing additional students.
06-May-15	Main industry partner's head office	Face-to-face meeting	Meeting with main industry partner contact and database manager to discuss the data.
07-May-15	N.A.	E-mail	Data meeting follow-up and further discussions focussed on their data.
28-May-15	N.A.	E-mail	Provided main industry partner with initial findings in a report related to the data provided thus far.
01-Jun-15	N.A.	E-mail	Received first feedback on report of 28 May 2016.
21-Jul-15	N.A.	E-mail	Discussion related to data cleaning.
28-Jul-15	N.A.	E-mail	Requested blank bursary contracts and application forms from main industry partner.
30-Jul-15	N.A.	E-mail	Received blank contracts, blank application forms and other information related to the application process of the main industry partner.
05-Aug-15	N.A.	E-mail	Correspondence related to data cleaning.
13-Aug-15	N.A.	E-mail	Received third sample of data, containing all students currently in the main industry partner's database.
21-Aug-15	N.A.	E-mail	Organised time for meeting with main industry partner management and consultant.
25-Aug-15	Main industry partner's head office	Face-to-face meeting	Two meetings. The first with the main industry partner contact and other data managers to discuss specific data collection methods used by them, their application process, and data cleaning. The second meeting was held with the main industry partner contact, other managers, and a consultant they have hired to re-engineer their systems. The second meeting was to discuss the strategic review of the NGO and how the work of this thesis study may be used to assist them.
27-Aug-15	N.A.	E-mail	Discussion about data cleaning for Case Study A.
05-Oct-15	N.A.	E-mail	Follow up after meeting of 25 August 2016. Also sent through informal report as was requested by main industry partner during the same meeting.
06-Oct-15	Stellenbosch University	Face-to-face meeting	Discussion with industry expert around proposed research methodology in general.
13-Oct-15	N.A.	E-mail	First contact made with secondary industry partner, requesting sample data.

20-Oct-15	N.A.	E-mail	Main industry partner made specific requests for Case Study A.
21-Oct-15 – 26-Oct-15	N.A.	E-mail	Discussion about data cleaning for Case Study A.
29-Oct-15	N.A.	E-mail	Secondary industry partner agreed to provide data and signs NDA.
02-Nov-15 – 11-Nov-15	N.A.	E-mail	Discussion about data cleaning for Case Study A.
16-Nov-15	N.A.	E-mail	Received articles from main industry partner related to study and #FeesMustFall.
23-Nov-15	N.A.	E-mail	Provided a list of questions related to data sent through to main industry partner.
26-Nov-15	N.A.	E-mail	Received sample data from secondary industry partner (for Case Study C).
07-Dec-15	N.A.	E-mail	Received answers from main industry partner on questions related to data.
19-Jan-16	N.A.	E-mail	Sent specific statistics as requested by main industry partner.
26-Jan-16	N.A.	E-mail	Send data cleaning chapter of thesis (Chapter 7) to verify correctness with main industry partner.
27-Jan-16	N.A.	E-mail	Received feedback on data cleaning chapter (for Case Study A).
01-Feb-16	N.A.	E-mail	Sent report requested by main industry partner focussed on score fields (for Case Study A).
02-Feb-16	N.A.	E-mail	Questions and suggestions by main industry partner based on report.
08-Feb-16	N.A.	E-mail	Updated and final report sent to main industry partner.
09-Feb-16 – 10-Feb-16	N.A.	E-mail	Received initial feedback on report from main industry partner and provide further clarification on certain issues for main industry partner.
11-Mar-16	Stellenbosch University	Presentation and discussion	Meeting with main industry partner to discuss and approve final use case and activity diagrams. Also to present progress of thesis work and further discuss the functionality of the proposed DSS concept demonstrator.
14-Feb-16	N.A.	E-mail	Meeting follow-up and discussion on final data sample to be received from main industry partner (for Case Study B).
16-Feb-16 – 18-Feb-16	N.A.	E-mail	Discussion on data cleaning for Case Study B.
07-Mar-16	N.A.	E-mail	Discussion about data cleaning for Case Study B.
18-Mar-16	N.A.	E-mail	Obtained final data sample from main industry partner (for Case Study B).
15-Apr-16	Stellenbosch University	Face-to-face meeting	Discussion with industry expert around proposed research methodology in general.

18-Apr-16	N.A.	E-mail	Organised time for meeting with main industry partner management to present thesis progress and user interface prototype.
05-May-16	Main industry partner's head office	Presentation and discussion	Presented study progress and preliminary user interface to main industry partner's management and discuss its functionality. Received valuable feedback.
06-Jul-16	N.A.	E-mail	Main industry partner sent their annual report containing a brief summary of the current study in collaboration with them using their data.
06-Jul-16	N.A.	E-mail	Main industry partner provided trends based on past data with respect to the number of students who applied and were successful in receiving a bursary.
21-Jul-16	N.A.	E-mail	Fifteen-minute video demonstration sent to main industry partner so as to obtain more feedback and enquire on their desire to implement it.
21-Jul-16	N.A.	E-mail	Fifteen-minute video demonstration sent to other members of industry so as to obtain feedback on whether they believe such a system is feasible.
28-Jul-16	N.A.	E-mail	Sample data obtained from secondary main industry partner.
04-Aug-16	Office of database programmers employed by the main industry partner.	Face-to-face meeting	Based on the video demonstration sent to the main industry partner on 21 July 2016 they expressed a desire to have a similar system integrated on their existing database. A meeting was thus set up with their database programmers to discussed its feasibility. The purpose of the meeting was to first present a brief demonstration of the proposed DSS and its workings, followed by a discussion on its components using supporting documentation. Based on this, the remainder of the discussion centred around whether it is within the capabilities of the external company to design the system for the main industry partner.
18-Aug-16	N.A.	E-mail	Database programmers employed by main industry partner provided feedback on their thoughts on the feasibility of the proposed DSS.
23-Aug-16	N.A.	E-mail	Initial variable importance analysis report of the sample data of Case Study B sent to main industry partner.
26-Aug-16	N.A.	E-mail	Main industry partner provided feedback on the initial Case Study B variable importance analysis report.

APPENDIX B

R libraries employed

The specific libraries, and functions of the libraries as well as other important information related to these functions, that were employed in this study may be found in Table B.1.

TABLE B.1: *Main R libraries and functions used in study and during creation of the DSS.*

Purpose	Related section in literature study	Library	Function	Reference for function	Other important settings related to the function
Base models					
Logistic regression	§3.4	stats	glm	[55], [136], [238]	family = “binomial”, predictive cut-off = 0.5
Classification and regression trees	§3.6	rpart	rpart	[29]	control = rpart.control(cp = 0, xval = 10), parms = list(split = “gini”),
	§3.6		prune	[29]	cp = bestcp
Random forests	§3.7	randomForest	randomForest	[31]	replace=TRUE, ntree=15000, mtry = round(sqrt(length(IndVars))
The C4.5 algorithm	§3.8	RWeka	J48	[182], [184]	Weka_control(R = TRUE, M = 2)
Support vector machines	§3.9	kernlab	ksvm	[46], [123], [179] [36]	kernel = “rbfdot”, C = 1, type = “C-svc” kpar = “automatic”
Assumptions					
Independence of residuals (Durbin-Watson)	§3.4 §3.10.2 §3.10.2	stats car car	glm durbinWatsonTest durbinWatsonTest	[55], [136], [238] [68] [68]	family = “binomial”, predictive cut-off = 0.5 -
Measurement errors	§3.2.1, §3.10.1	performed manually	-	-	-
Limited multicollinearity (VIF)	§3.10.2	rms	vif	[49]	-
Detecting outliers (Random forests outlier detection)	§2.2.4, §3.10.2	CORElearn	CoreModel rfOutliers	[30], [187] [30]	model = c(“rf”) -
Detecting influential observations (Cook’s Distance)	§2.2.4, §3.10.2	stats	cooks.distance	[44]	-
Mutually exclusive categorical dependent variables	§3.10.3	performed manually	-	-	-
Minimum sample size	§3.10.3, §3.10.4, §3.10.5, §3.10.6, §3.10.7	performed manually	-	-	-
Correct coding of dependent variable	§3.10.3	performed manually	-	-	-
Linearity of the logit (Box-Tidwell)	§3.10.3	stats, performed manually	glm	[55], [136]	family = “binomial”, significance cut-off = 0.05
Ensemble					
Base model predictability weighting	§4.2.1	performed manually	-	-	-
Combined prediction of base models	§4.2.1	performed manually	-	-	-
MCDA outranking method (ELECTRE III)	§4.3.5	OutrankingTools	Electre3_Simple-Thresholds	[191]	Refer to §4.3.5 for threshold selection
Model fit and variable importance					
Log likelihood	§3.5.1	epiDisplay	lrtest	-	-
p-values, $\hat{\beta}$s, OR, CIs	§3.5.2	stats, manually	confint, glm	[55], [136]	level = 0.95, α = 5%
RF variable importance	§3.7, §3.7.1	randomForest	varImpPlot,	[31]	importance = TRUE
CART	§3.6	rpart.plot	prp	[147]	type = 4, extra = 104
Testing model predictive accuracy	§3.5.3	performed manually	-	-	-
Other					
Formatting of tables and graphs	-	formattable, ggplot2	-	-	-
Data cleaning, Data splitting	§2.2.4, §3.1.11	performed manually	-	-	set.seed(sample(1:10000, 1))
Creation of GUI	§2.3.2	Shiny	-	-	-

APPENDIX C

Additional graphs and tables of Case Study B

Frequency bar plots were drawn up for the sample data of Case study B (§8.3) displaying the percentages of students who graduated successfully as a function of specific Grade 12 subject marks. The plots for the independent variables *Grade 12 Afrikaans home language*, *Grade 12 Afrikaans second language*, *Grade 12 Agricultural science*, *Grade 12 Business economics*, *Grade 12 Economics*, *Grade 12 Geography*, *Grade 12 History*, *Grade 12 isiXhosa home language*, *Grade 12 isiZulu home language*, *Grade 12 Life sciences*, *Grade 12 Mathematics literacy*, and *Grade 12 Setswana home language* are shown in Figures C.1–C.12.

Tables were also drawn up for the sample data of Case Study B (§8.3) displaying the proportion of students within each percentage class of specific high school subject marks or specific high school average marks *versus* their attendance and graduation rate of specific *Tertiary institutions*, *Study fields*, *Qualification types*, and *Study regions*. Results of this kind for *Grade 11 November average*, *Grade 12 Mathematics*, and *Grade 12 Physical science* are shown in Tables C.1–C.12

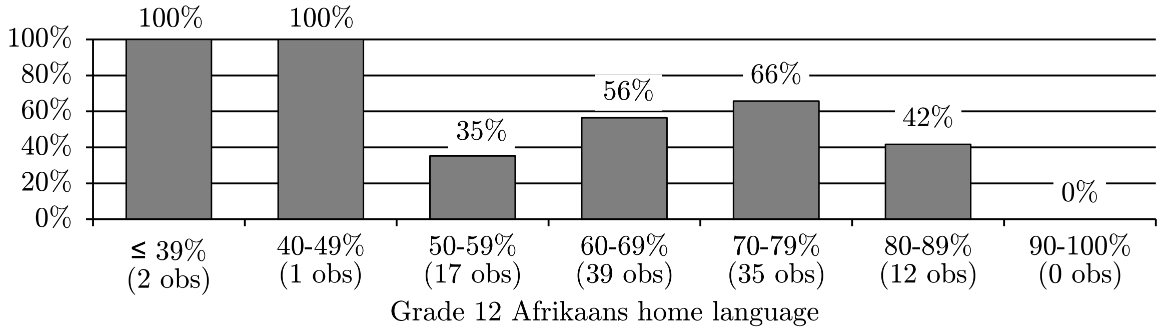


Figure C.1: *Percentages of students who graduated successfully as a function of Grade 12 Afrikaans home language mark in the sample of Case Study B.*

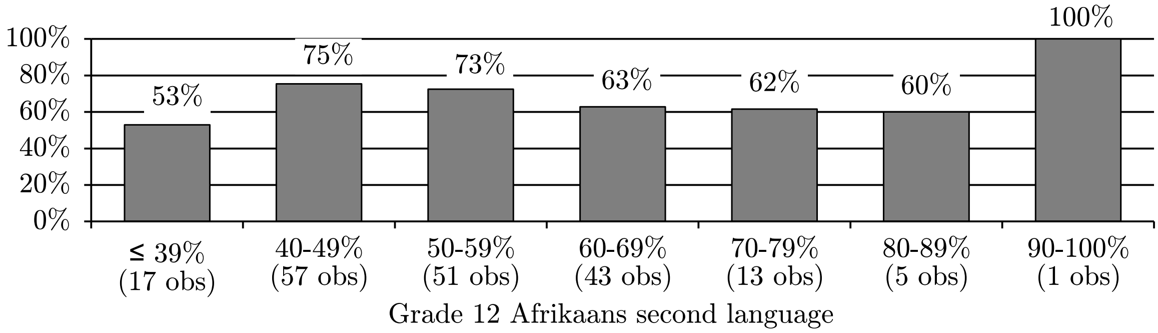


FIGURE C.2: *Percentages of students who graduated successfully as a function of Grade 12 Afrikaans second language mark in the sample of Case Study B.*

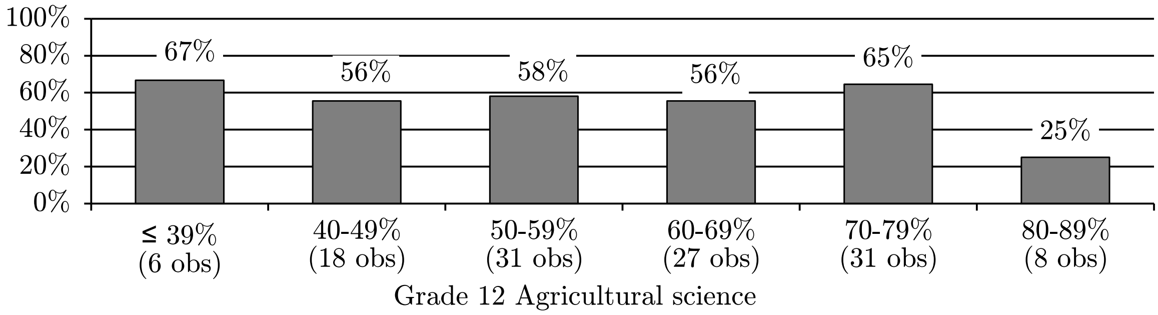


FIGURE C.3: *Percentages of students who graduated successfully as a function of Grade 12 Agricultural science mark in the sample of Case Study B.*

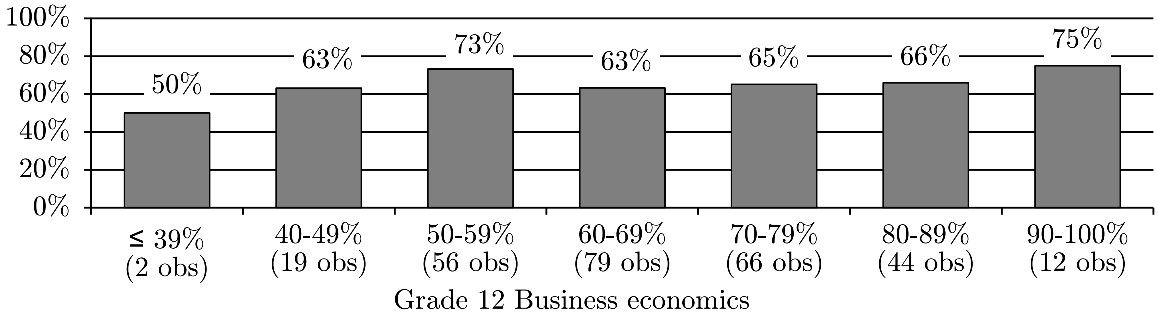


FIGURE C.4: *Percentages of students who graduated successfully as a function of Grade 12 Business economics mark in the sample of Case Study B.*

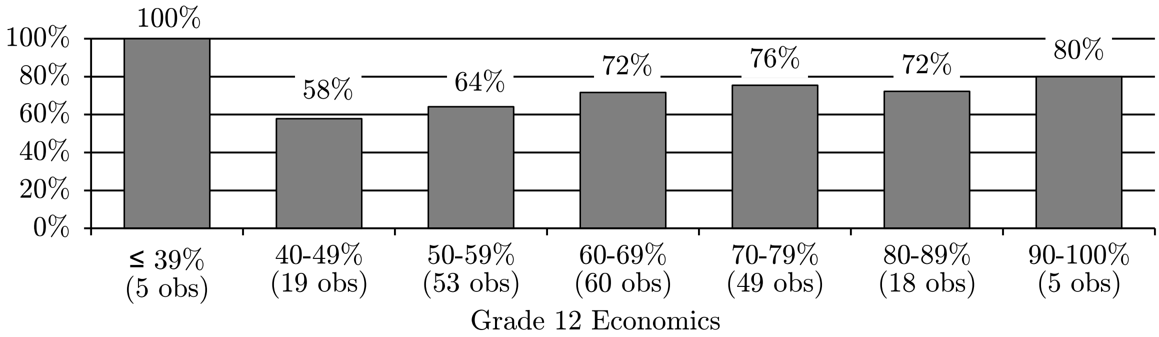


FIGURE C.5: *Percentages of students who graduated successfully as a function of Grade 12 Economics mark in the sample of Case Study B.*

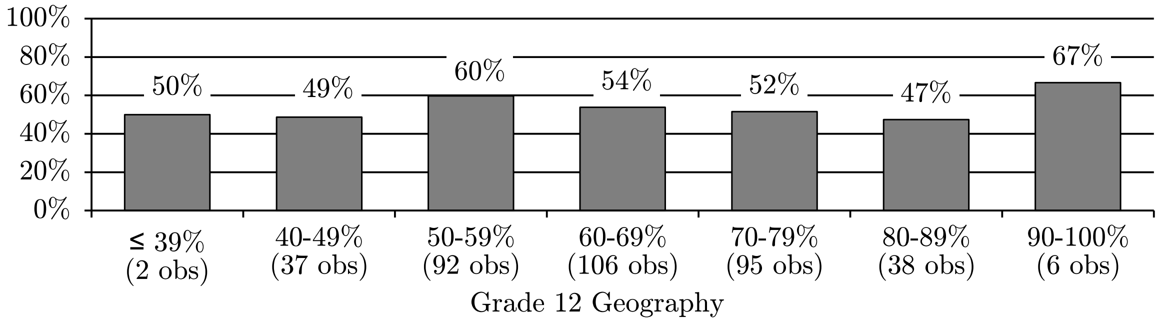


Figure C.6: *Percentages of students who graduated successfully as a function of Grade 12 Geography mark in the sample of Case Study B.*

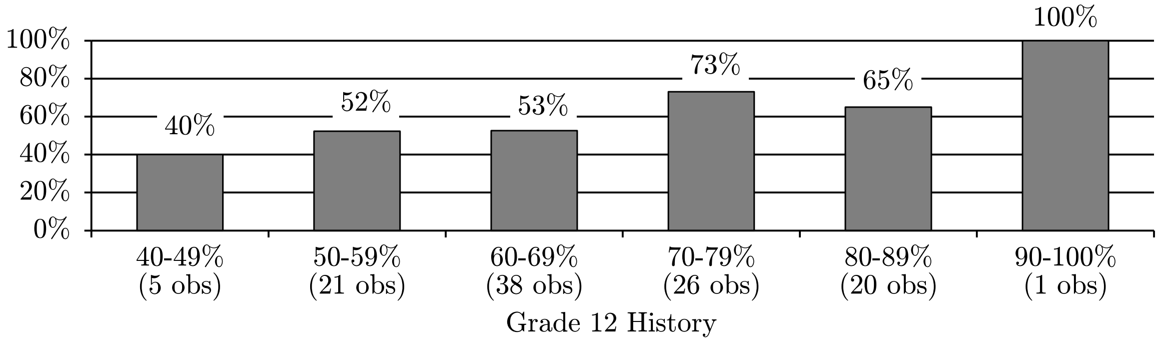


Figure C.7: *Percentages of students who graduated successfully as a function of Grade 12 History mark in the sample of Case Study B.*

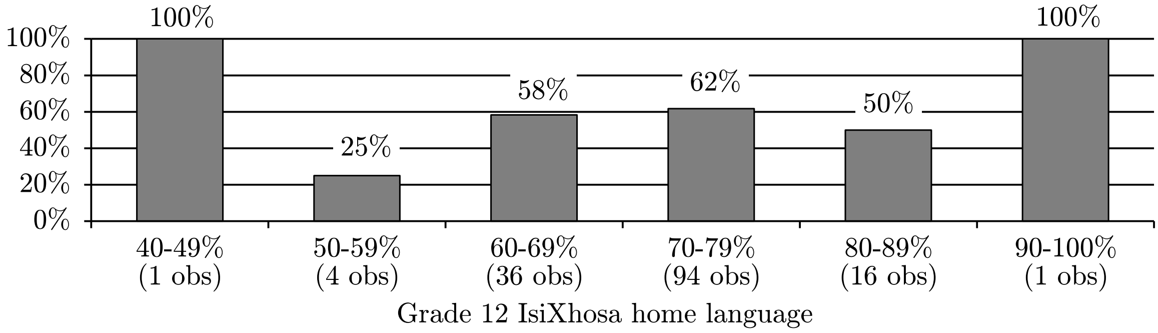


Figure C.8: *Percentages of students who graduated successfully as a function of Grade 12 isiXhosa home language mark in the sample of Case Study B.*

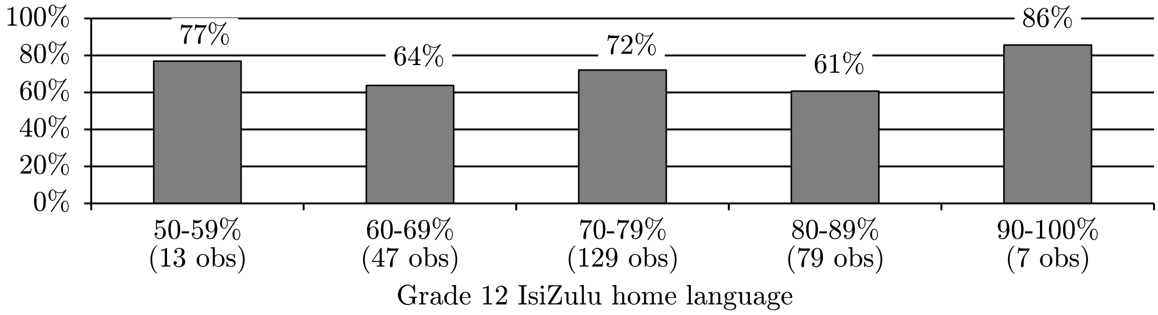


Figure C.9: *Percentages of students who graduated successfully as a function of Grade 12 isiZulu home language mark in the sample of Case Study B.*

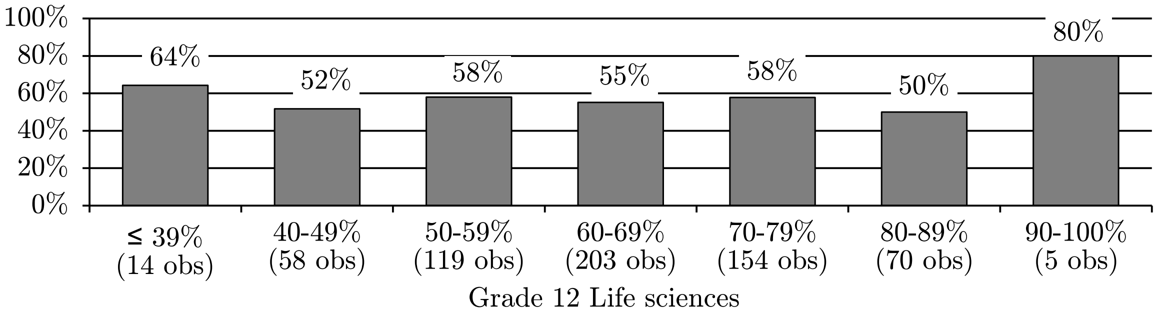


FIGURE C.10: *Percentages of students who graduated successfully as a function of Grade 12 Life science mark in the sample of Case Study B.*

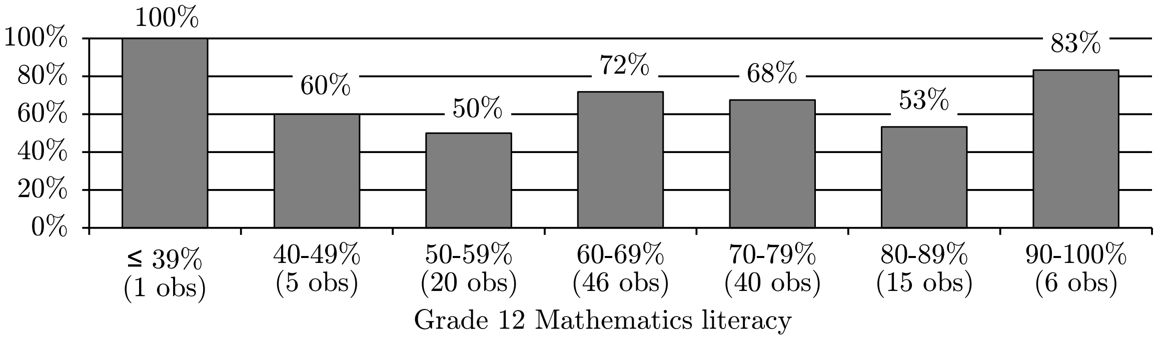


FIGURE C.11: *Percentages of students who graduated successfully as a function of Grade 12 Mathematics literacy mark in the sample of Case Study B.*

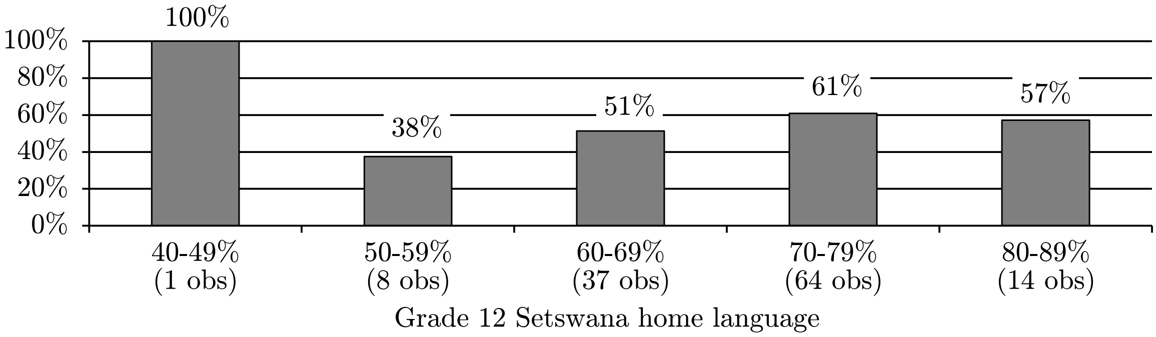


FIGURE C.12: *Percentages of students who graduated successfully as a function of Grade 12 Setswana home language mark in the sample of Case Study B.*

Table C.1: *Proportions of students within each percentage class of Grade 11 November marks who attended a tertiary institution and graduated at a tertiary institution, respectively.*

	Proportions of students within each % class of Grade 11 Nov marks who:									
	attended a tertiary institution					graduated at a tertiary institution				
	40–49	50–59	60–69	70–79	80–89	40–49	50–59	60–69	70–79	80–89
SU	0%	0%	2%	6%	8%	☒	☒	0%	20%	100%
UP	0%	1%	5%	10%	8%	☒	0%	14%	0%	0%
Rhodes	0%	6%	2%	1%	0%	☒	33%	0%	0%	☒
Wits	0%	7%	12%	19%	8%	☒	57%	24%	7%	0%
UJ	0%	1%	12%	17%	17%	☒	0%	47%	38%	0%
UCT	0%	2%	5%	4%	8%	☒	50%	43%	33%	100%
UFS	13%	7%	1%	0%	8%	0%	43%	100%	☒	100%
UWC	13%	5%	6%	8%	0%	0%	40%	33%	50%	☒
CUT	13%	4%	4%	3%	0%	100%	75%	67%	100%	☒
NMMU	13%	15%	7%	3%	0%	100%	64%	60%	50%	☒
NWU	0%	4%	5%	0%	0%	☒	75%	43%	☒	☒
UKZN	25%	18%	21%	18%	33%	100%	65%	65%	86%	0%
SMHSU	0%	0%	1%	1%	0%	☒	☒	100%	100%	☒
TUT	0%	3%	2%	1%	0%	☒	33%	100%	100%	☒
DUT	25%	18%	9%	5%	8%	100%	59%	100%	75%	100%
CPUT	0%	7%	7%	3%	0%	☒	86%	73%	100%	☒
# obs	8	95	147	77	12	6	55	79	33	4

Table C.2: *Proportions of students within each percentage class of Grade 11 November marks who studied in a study field and graduated within a specific field of study, respectively.*

	Proportions of students within each % class of Gr.11 Nov marks who:									
	studied in a study field					graduated in a study field				
	40–49	50–59	60–69	70–79	80–89	40–49	50–59	60–69	70–79	80–89
Eng	11%	15%	16%	29%	25%	100%	43%	17%	27%	0%
Law	11%	4%	10%	4%	0%	0%	0%	36%	33%	☒
Sci	22%	20%	24%	32%	25%	0%	63%	40%	36%	33%
Tech	0%	4%	3%	3%	8%	☒	100%	25%	50%	100%
Comm	0%	17%	14%	13%	25%	☒	44%	67%	60%	33%
Blt Env	0%	1%	1%	3%	0%	☒	100%	50%	50%	☒
Arts	0%	4%	3%	3%	8%	☒	75%	75%	50%	100%
Med	0%	7%	7%	4%	8%	☒	43%	70%	67%	0%
Hum	22%	12%	11%	9%	0%	100%	64%	94%	71%	☒
Edu	0%	4%	5%	1%	0%	☒	75%	75%	100%	☒
Man	33%	11%	6%	0%	0%	100%	80%	89%	☒	☒
Bus Man	0%	1%	1%	0%	0%	☒	100%	100%	☒	☒
# obs	9	95	147	77	12	6	55	79	33	4

TABLE C.3: *Proportions of students within each percentage class of Grade 11 November marks who pursued a specific qualification type and obtained a specific qualification type, respectively.*

	Proportions of students within each % class of Gr.11 Nov marks who:									
	pursued a specific qualification type					obtained a specific qualification type				
	40–49	50–59	60–69	70–79	80–89	40–49	50–59	60–69	70–79	80–89
Ext Deg	13%	4%	5%	5%	0%	0%	25%	43%	50%	╳
Degree	50%	56%	73%	78%	92%	50%	51%	47%	35%	27%
Nat Dip	38%	40%	22%	17%	8%	100%	71%	84%	77%	100%

TABLE C.4: *Proportions of students within each percentage class of Grade 11 November marks who studied in a specific study region and graduated in a specific study region, respectively.*

	Proportions of students within each % class of Gr.11 Nov marks who:									
	studied in a specific study region					graduated in a specific study region				
	40–49	50–59	60–69	70–79	80–89	40–49	50–59	60–69	70–79	80–89
GP	0%	13%	31%	49%	33%	╳	42%	38%	21%	0%
EC	11%	21%	9%	4%	0%	100%	55%	46%	33%	╳
FS	22%	12%	5%	3%	8%	50%	55%	71%	100%	100%
WC	11%	15%	20%	21%	17%	0%	64%	47%	44%	100%
LP	11%	0%	1%	0%	0%	0%	╳	100%	╳	╳
NW	0%	4%	5%	0%	0%	╳	75%	43%	╳	╳
KZN	44%	36%	30%	23%	42%	100%	62%	75%	83%	20%

TABLE C.5: *Proportions of students within each percentage class of Grade 12 Mathematics marks who attended a specific tertiary institution and graduated at a specific tertiary institution, respectively.*

	Proportions of students within each % class of Gr.12 Mathematics marks who:												
	attended a tertiary institution							graduated at a tertiary institution					
	40–49	50–59	60–69	70–79	80–89	≥90		40–49	50–59	60–69	70–79	80–89	≥90
SU	4%	1%	5%	2%	5%	11%		67%	0%	30%	50%	0%	0%
UP	7%	3%	5%	11%	6%	2%		60%	25%	10%	22%	0%	0%
Rhodes	1%	4%	5%	4%	1%	2%		100%	20%	22%	50%	0%	100%
Wits	1%	4%	10%	18%	16%	35%		100%	20%	39%	28%	44%	31%
UJ	9%	11%	15%	21%	17%	13%		17%	23%	38%	48%	71%	67%
UCT	4%	0%	1%	5%	8%	9%		67%	╳	100%	56%	25%	75%
UFS	8%	4%	1%	5%	5%	2%		83%	80%	0%	38%	40%	100%
UWC	1%	10%	5%	3%	3%	4%		100%	42%	100%	60%	67%	50%
CUT	1%	4%	4%	2%	6%	2%		100%	80%	14%	67%	33%	100%
NMMU	7%	7%	7%	4%	3%	0%		100%	67%	50%	67%	33%	╳
NWU	0%	3%	3%	2%	2%	3%	2%		75%	100%	67%	33%	100%
UKZN	19%	18%	14%	9%	14%	11%		62%	76%	58%	50%	50%	60%
SMHSU	1%	0%	2%	1%	1%	0%		100%	╳	100%	100%	100%	╳
TUT	5%	3%	4%	2%	6%	0%		75%	100%	88%	100%	83%	╳
DUT	14%	15%	13%	7%	4%	0%		70%	84%	74%	82%	75%	╳
CPUT	10%	8%	7%	3%	0%	7%		100%	90%	67%	100%	╳	100%
# obs	69	117	177	159	98	46		50	73	92	78	45	23

Table C.6: *Proportion students within each percentage class of Grade 12 Mathematics marks who studied in a specific study field and graduated in specific a study field, respectively.*

	Proportions of students within each % class of Gr.12 Mathematics marks who:											
	studied in a specific study field						graduated in specific a study field					
	40–49	50–59	60–69	70–79	80–89	≥90	40–49	50–59	60–69	70–79	80–89	≥90
Eng	7%	13%	25%	32%	39%	48%	40%	38%	39%	44%	34%	36%
Law	1%	3%	1%	0%	1%	2%	100%	25%	100%	╳	100%	0%
Sci	23%	24%	22%	23%	27%	26%	59%	53%	58%	34%	46%	67%
Tech	3%	6%	4%	5%	5%	4%	100%	57%	29%	56%	80%	50%
Comm	27%	26%	19%	21%	16%	15%	60%	75%	43%	59%	47%	71%
Blt Env	0%	2%	4%	3%	1%	0%	╳	0%	57%	60%	100%	╳
Arts	3%	2%	1%	0%	0%	0%	100%	67%	100%	╳	╳	╳
Med	8%	6%	12%	12%	8%	4%	67%	88%	68%	68%	75%	50%
Hum	8%	5%	6%	2%	0%	0%	100%	83%	55%	33%	╳	╳
Edu	5%	5%	4%	1%	1%	0%	100%	83%	63%	100%	0%	╳
Man	11%	5%	2%	1%	2%	0%	88%	83%	100%	100%	50%	╳
Bus Man	4%	3%	0%	0%	0%	0%	100%	75%	╳	╳	╳	╳
# obs	74	124	183	164	104	46	53	78	95	82	48	23

Table C.7: *Proportions of students within each percentage class of Grade 12 Mathematics marks who pursued a specific qualification type, obtained a specific qualification type, respectively.*

	Proportions of students within each % class of Gr.12 Mathematics marks who:											
	pursued a specific qualification type						obtained a specific qualification type					
	40–49	50–59	60–69	70–79	80–89	≥90	40–49	50–59	60–69	70–79	80–89	≥90
Ext Deg	14%	5%	3%	1%	3%	0%	60%	33%	50%	50%	0%	╳
Degree	42%	55%	58%	73%	68%	83%	68%	54%	44%	39%	41%	42%
Nat Dip	44%	40%	38%	26%	29%	17%	78%	81%	65%	81%	60%	88%
# obs	73	121	180	164	103	46	52	77	94	82	47	23

Table C.8: *Proportions of students within each percentage class of Grade 12 Mathematics marks who studied in a specific study region and graduated in a specific study region, respectively.*

	Proportions of students within each % class of Gr.12 Mathematics marks who:											
	studied in a specific study region						graduated in a specific study region					
	40–49	50–59	60–69	70–79	80–89	≥90	40–49	50–59	60–69	70–79	80–89	≥90
GP	23%	23%	35%	52%	47%	53%	53%	36%	42%	38%	54%	42%
EC	9%	11%	12%	7%	5%	2%	100%	50%	41%	58%	40%	100%
FS	11%	10%	6%	8%	13%	4%	75%	83%	18%	54%	38%	100%
WC	19%	19%	17%	14%	16%	29%	86%	61%	65%	64%	25%	46%
LP	4%	1%	2%	1%	0%	0%	67%	0%	67%	50%	╳	╳
NW	3%	2%	2%	2%	2%	0%	100%	100%	100%	67%	50%	╳
KZN	31%	33%	26%	16%	18%	11%	65%	80%	66%	64%	56%	60%
# obs	74	121	179	161	102	45	53	76	93	79	48	22

TABLE C.9: *Proportions of students within each percentage class of Grade 12 Physical science marks who studied in a specific study region and graduated in a specific tertiary institution, respectively.*

	Proportions of students within each % class of Gr.12 Physical science marks who:											
	attended a tertiary institution						graduated at a tertiary institution					
	≤ 39	40–49	50–59	60–69	70–79	80–89	≤ 39	40–49	50–59	60–69	70–79	80–89
SU	11%	6%	3%	3%	6%	0%	0%	33%	25%	20%	43%	☒
UP	6%	10%	7%	6%	8%	14%	100%	20%	30%	11%	11%	17%
Rhodes	0%	2%	6%	3%	0%	5%	☒	0%	50%	20%	☒	50%
Wits	0%	4%	7%	12%	26%	30%	☒	50%	30%	50%	30%	31%
UJ	11%	4%	20%	20%	11%	12%	50%	50%	37%	53%	46%	20%
UCT	0%	6%	3%	3%	6%	9%	☒	33%	75%	50%	57%	25%
UFS	6%	4%	4%	4%	1%	2%	100%	50%	80%	33%	0%	0%
UWC	17%	10%	6%	6%	3%	0%	0%	80%	63%	67%	50%	☒
CUT	6%	6%	3%	7%	2%	2%	100%	33%	50%	36%	50%	100%
NMMU	6%	8%	3%	7%	3%	7%	100%	75%	50%	45%	50%	67%
NWU	0%	4%	1%	3%	1%	0%	☒	100%	50%	80%	100%	☒
UKZN	11%	12%	12%	9%	19%	7%	100%	67%	69%	54%	64%	0%
SMHSU	6%	2%	1%	1%	1%	2%	100%	100%	100%	100%	100%	100%
TUT	0%	12%	4%	5%	3%	0%	☒	83%	80%	100%	67%	☒
DUT	22%	12%	10%	7%	8%	5%	100%	83%	57%	70%	78%	50%
CPUT	0%	2%	11%	2%	2%	5%	☒	100%	93%	67%	100%	100%
# obs	18	52	137	148	115	43	12	32	76	76	55	15

TABLE C.10: *Proportion students within each percentage class of Grade 12 Physical science marks who studied in a specific study field and graduated in specific a study field, respectively.*

	Proportions of students within each % class of Gr.12 Physical science marks who:											
	studied in a specific study field						graduated in specific a study field					
	≤ 39	40–49	50–59	60–69	70–79	80–89	≤ 39	40–49	50–59	60–69	70–79	80–89
Eng	0%	23%	23%	42%	38%	53%	☒	50%	47%	42%	31%	35%
Law	5%	0%	1%	1%	2%	0%	0%	☒	50%	100%	50%	☒
Sci	19%	25%	32%	33%	29%	30%	50%	33%	53%	61%	44%	15%
Tech	0%	7%	5%	3%	7%	5%	☒	75%	57%	50%	63%	100%
Comm	24%	7%	5%	6%	8%	0%	80%	50%	43%	50%	78%	☒
Blt Env	5%	2%	4%	2%	2%	0%	100%	100%	40%	33%	50%	☒
Arts	19%	0%	1%	0%	1%	0%	75%	☒	100%	☒	100%	☒
Med	0%	15%	18%	9%	8%	12%	☒	100%	72%	57%	78%	60%
Hum	10%	12%	7%	0%	4%	0%	50%	86%	60%	☒	60%	☒
Edu	0%	3%	3%	2%	2%	0%	☒	50%	50%	67%	100%	☒
Man	19%	7%	1%	2%	0%	0%	100%	75%	100%	100%	☒	☒
Bus Man	0%	0%	1%	0%	0%	0%	☒	☒	0%	☒	☒	☒
# obs	21	60	141	154	117	43	15	37	78	81	56	15

Table C.11: *Proportions of students within each percentage class of Grade 12 Physical science marks who pursued a specific qualification type, obtained a specific qualification type, respectively.*

	Proportions of students within each % class of Gr.12 Physical science marks who:											
	pursued a specific qualification type						obtained a specific qualification type					
	≤ 39	40–49	50–59	60–69	70–79	80–89	≤ 39	40–49	50–59	60–69	70–79	80–89
Ext Deg	0%	5%	4%	5%	3%	7%	☒	33%	67%	50%	0%	0%
Degree	52%	53%	58%	58%	73%	72%	55%	53%	44%	45%	42%	29%
Nat Dip	48%	42%	38%	37%	24%	21%	90%	71%	72%	65%	71%	67%
# obs	21	57	140	154	116	43	15	34	78	81	56	15

Table C.12: *Proportions of students within each percentage class of Grade 12 Physical science marks who studied in a specific study region and graduated in a specific study region, respectively.*

	Proportions of students within each % class of Gr.12 Physical science marks who:											
	studied in a specific study region						graduated in a specific study region					
	≤ 39	40–49	50–59	60–69	70–79	80–89	≤ 39	40–49	50–59	60–69	70–79	80–89
GP	16%	26%	37%	44%	48%	58%	67%	53%	38%	54%	34%	28%
EC	11%	9%	9%	12%	3%	12%	100%	60%	50%	44%	50%	60%
FS	11%	16%	8%	12%	3%	5%	100%	56%	73%	39%	33%	50%
WC	26%	21%	22%	14%	17%	14%	0%	58%	74%	52%	55%	50%
LP	5%	4%	1%	1%	2%	0%	100%	50%	50%	100%	50%	☒
NW	0%	4%	1%	3%	1%	0%	☒	100%	100%	80%	100%	☒
KZN	32%	21%	22%	15%	26%	12%	100%	75%	63%	61%	68%	20%
# obs	19	57	139	153	117	43	13	35	78	81	56	15