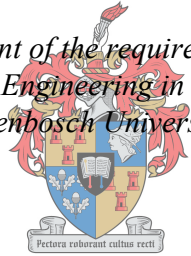


DATA MINING CONSTRUCTION PROJECT INFORMATION TO AID PROJECT MANAGEMENT

by
Louis Johan Botha

*Thesis presented in fulfilment of the requirements for the degree of
Master of Engineering in Civil Engineering in the Faculty of Engineering
at Stellenbosch University*



UNIVERSITEIT
iYUNIVESITHI
STELLENBOSCH
UNIVERSITY

100
1918 · 2018

Supervisor: Prof Jan Wium

December 2018

Declaration:

By submitting this thesis electronically, I declare that the entirety of the work contained herein is my own, original work, that I am the sole author therefor (save to the extent explicitly otherwise stated), the 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.

December 2018

Louis Johan Botha

Copyright © 2018 Stellenbosch University
All rights reserved

Abstract:

Internationally, the popularity of data mining and its use in a business context has grown rapidly in many sectors. The organisations that utilise data mining have experienced significant gains in efficiency, productivity and profitability. The utilisation of data mining within the construction industry has however lagged behind other sectors, especially in South Africa.

Data mining to aid project management has seen limited application. The leader in applying data mining to improve project management has been the software development sector as it is plagued by project cost and time overruns and a high number of failed projects. The construction industry in South Africa suffers from similar cost and time overruns, yet data mining in the construction sector has been limited. Few applications exist of data mining to improve the management of construction projects.

The process followed to implement a data mining application has been largely focused on the specific statistical and technical details of the data preparation and the data mining model. These details are inherently application specific and do not provide a general data mining process. Guides that define and demonstrate the general data mining process are limited or outdated, with no such guide existing for data mining in the construction sector.

The research examines the application of data mining to the construction sector and to the improvement of project management in the software development sector. From these sources and a discussion of construction projects in South Africa, a comprehensive data mining process is synthesised. The data mining process is discussed in the context of the construction sector in South Africa and construction sector personnel with limited experience of data mining. A number of user-friendly, yet rigorous, data mining resources are presented. A selection of these resources are applied to a real project dataset obtained from the Western Cape Government's Department of Public Works' internal project database. A data mining application is developed by adhering to the data mining process defined within the research. The results were discussed along with several salient lessons learned.

Opsomming:

Internasionaal het die gewildheid van data-ontginning en die gebruik daarvan in 'n besigheidskonteks vinnig in baie sektore gegroei. Die organisasies wat data-ontginning gebruik, het aansienlike winste in doeltreffendheid, produktiwiteit en winsgewendheid ervaar. Die gebruik van data-ontginning in die konstruksiebedryf het veral in Suid-Afrika agterweë gebly.

Data-ontginning om projekbestuur te help, het beperkte toepassing gesien. Die leier in die toepassing van data-ontginning om projekbestuur te verbeter, was die sagteware-ontwikkelingsektor aangesien dit gepla word deur projek koste en tydoorskrydings en 'n groot aantal mislukte projekte. Die konstruksiebedryf in Suid-Afrika ly aan soortgelyke koste- en tydoorskrydings, maar data-ontginning in die konstruksiesektor is beperk. Min toepassings van data-ontginning om die bestuur van konstruksieprojekte te verbeter, bestaan.

Die proses wat gevolg is om 'n data-ontginnings aansoek te implementeer, het hoofsaaklik gefokus op die spesifieke statistiese en tegniese besonderhede van die data-voorbereiding en die data-ontginningsmodel. Hierdie inligting is inherent toepassingspesifiek en is geneig om af te sien daarvan om algemene advies te gee. Gidse wat die algemene data-ontginningsproses definieer en demonstreer is beperk of verouderd en geen sodanige gidse vir data-ontginning in die konstruksiesektor bestaan nie.

Die navorsing ondersoek die toepassing van data-ontginning aan die konstruksiesektor en die verbetering van projekbestuur in die sagteware-ontwikkelingsektor. Uit hierdie bronne en 'n bespreking van konstruksieprojekte in Suid-Afrika is 'n omvattende data-ontginningsproses gesintetiseer. Die data-ontginningsproses is bespreek in die konteks van die konstruksiesektor in Suid-Afrika en konstruksiesektorpersoneel met beperkte ervaring van data-ontginning. 'n Aantal gebruikersvriendelike dog streng data-ontginningsbronne is aangebied. 'n Seleksie van hierdie hulpbronne is toegepas op 'n werklike projekdatastel wat verkry is van die Wes-Kaapse regering se Departement van Openbare Werke se interne projekdatabasis. 'n Data-ontginningstoepassing is ontwikkel deur te voldoen aan die data-ontginningsproses wat binne die navorsing gedefinieer is. Die uitslae is bespreek met verskeie belangrike lesse wat geleer is.

Acknowledgements:

This research would not have been possible without the constant assistance and support I received. I would like to thank the following contributors:

- Prof Wium, my study leader, for his guidance and support in formulating and realising the research.
- The Directorate of Construction and Maintenance at the Western Cape Government Department of Transport and Public Works for the data provided to the research.
- SPPrac for funding my postgraduate degree.
- My family and loved ones for their constant support and encouragement.

Table of Contents:

Declaration:.....	i
Abstract:.....	ii
Opsomming:.....	iii
Acknowledgements:.....	iv
Table of Contents:.....	v
Table of Figures:.....	xi
List of Tables:.....	xiii
1 Introduction:.....	1
1.1 Introduction and Background:.....	1
1.2 Research Aims and Objectives:.....	2
1.3 Scope and Limitations:.....	3
1.4 Methodology:.....	4
1.5 Research Outline:.....	7
1.5.1 Chapter 2: Literature Review:.....	8
1.5.2 Chapter 3: Data Mining Process:.....	8
1.5.3 Chapter 4: Data Mining Resources:.....	8
1.5.4 Chapter 5: Data Mining Implementation Example:.....	8
1.5.5 Chapter 6: Conclusion:.....	8
1.5.6 Chapter 7: Recommendations for Further Research:.....	8
2 Literature Review:.....	9
2.1 Introduction:.....	9
2.2 The Value of Data:.....	10
2.2.1 Value of Data and Data Driven Companies:.....	10

2.2.2	Impact of Poor Data:.....	10
2.3	Case Studies of Data Mining and Text Mining Used in Construction and Project Management:.....	12
2.3.1	Case Study 1: Residual Value Assessment of Heavy Construction Equipment: 12	
2.3.2	Case Study 2: Using Data Mining to Discover Knowledge in Enterprise Performance Records:.....	14
2.3.3	Case Study 3: Data Mining Tender Bid Information to Evaluate the Best Bid Selection Policy:.....	16
2.3.4	Case Study 4: Predicting Construction Cost Overruns Using Data Mining:.....	18
2.3.5	Case Study 5: Data Mining to Detect Early Warning Signs of Project Failure by Mining Unstructured Text from Site Meetings:	20
2.4	Case Studies of Data Mining Used to Improve Project Management in Software Development:	21
2.4.1	Case Study 6: Data Mining Application in a Software Project Management Process: 21	
2.4.2	Case Study 7: Data Mining for the Management of Software Development Processes.....	23
2.4.3	Case Study 8: Data Mining Applied to the Improvement of Project Management:	25
2.5	Construction Project Environment:.....	27
2.5.1	Construction Project Management Phases:.....	27
2.5.2	Sustainable Project Management Success Criteria:.....	29
2.6	Discussion of the Literature:	30
2.6.1	Summary of Construction Data Mining Case Studies:	31
2.6.2	Suitability of Applying Data Mining to Project Management in the Construction Sector: 32	
2.6.3	Definition of a Data Mining Process for Application in Construction Projects:35	
2.6.3.1	Knowledge from Case Studies with Defined Data Mining Processes:	35
2.6.3.2	Knowledge from Case Studies without a Defined Data Mining Process: ..	37

2.6.3.3	Data Mining Process:.....	38
2.7	Conclusion:.....	39
3	Data Mining Process:.....	40
3.1	Introduction and Background Information:.....	40
3.1.1	Types of Data:.....	42
3.1.2	Supervised vs Unsupervised Machine Learning:.....	44
3.1.3	Examples of Data Mining to Facilitate Construction Project Management:	45
3.2	Goal Definition:.....	46
3.2.1	Setting a Goal to Facilitate Construction Project Management Through Data Mining	46
3.2.2	Goal Definition Applied to Construction Project Management Examples:.....	47
3.3	Data Acquisition:.....	47
3.3.1	Information Extraction:.....	48
3.3.2	Typical Documentation in a South African Construction Project:	50
3.3.3	Building Information Models as a Source of Data:	52
3.3.4	Data Acquisition for Construction Project Management Examples	54
3.4	Pre-processing:	55
3.4.1	Goal Re-examination:	56
3.4.2	Feature Extraction and Selection:	56
3.4.3	Data Cleaning:	57
3.4.4	Data Reduction and Feature Transformation:.....	58
3.4.5	Pre-processing for Construction Project Management Examples:	60
3.5	Mining and Modelling:.....	61
3.5.1	Similarity and Distances:	62
3.5.2	Association Pattern Mining:	63
3.5.3	Cluster Analysis:.....	64

3.5.4	Outlier Analysis:	66
3.5.5	Data Classification and Regression:	67
3.5.5.1	Decision Trees:	69
3.5.5.2	Rule-Based Classifiers:.....	70
3.5.5.3	Probabilistic Classifiers:.....	71
3.5.5.4	Neural Networks:.....	72
3.5.5.5	<i>K</i> -Nearest-Neighbour:	73
3.5.5.6	Support Vector Machines:.....	73
3.5.5.7	Other Common Models:	75
3.5.6	Text Mining:	75
3.5.7	Other Data Mining Types:	76
3.5.8	Mining and Modelling Applied to Construction Project Management Examples 78	
3.6	Validation and Evaluation:.....	78
3.6.1	Cluster Validation:	78
3.6.1.1	Internal Cluster Validation Criteria:.....	79
3.6.1.2	External Cluster Validation Criteria:.....	80
3.6.1.3	Relative Validation Criteria:.....	80
3.6.2	Outlier Validation:	81
3.6.2.1	Internal Outlier Validation:	81
3.6.2.2	External Outlier Validation:	82
3.6.3	Classifier and Regression Evaluation:	83
3.6.3.1	Classifier Performance Measures:	83
3.6.3.2	Regression Performance Measures:.....	86
3.6.3.3	Classifier and Regression Evaluation:.....	87
3.6.4	Validation of Construction Project Examples.....	88

3.7	Conclusion:	89
4	Data Mining Resources:	91
4.1	Introduction:	91
4.2	Scikit-learn:	92
4.3	Natural Language Toolkit:	95
4.4	Machine Learning with R (mlr):	96
4.5	Orange:	97
4.6	RapidMiner:	98
4.7	Evaluation of Data Mining Resources:	99
4.8	Conclusion:	101
5	Data Mining Implementation Example:	102
5.1	Introduction:	102
5.2	Prediction of Employment Opportunities a Construction Project in South Africa will Create:	102
5.2.1	Goal Definition:	103
5.2.2	Data Acquisition:	103
5.2.3	Pre-Processing:	103
5.2.3.1	Goal Definition (Re-examination):	103
5.2.3.2	Feature Extraction and Selection, Data Cleaning, and Feature Transformation:	106
5.2.4	Mining and Modelling:	108
5.2.5	Validation and Evaluation:	108
5.2.6	Iterations:	110
5.3	Prediction of Project Cost Overruns:	111
5.3.1	Goal Definition:	111
5.3.2	Data Acquisition:	111
5.3.3	Pre-processing:	111

5.3.4	Discussion of Application Failure:	112
5.4	Lessons Learned from Applying the Data Mining Process:	113
5.5	Conclusion:.....	115
6	Conclusion:.....	116
6.1	Synthesis and Discussion of the Data Mining Process:	116
6.2	Data Mining Resources:	118
6.3	Data Mining Demonstration:.....	120
7	Recommendations for Further Research	122
8	References:	124
9	Appendix 1: Literature Review Summary:.....	128
10	Appendix 2: Chapter 3 Summary:	130
11	Appendix 3: Data Mining Resources Summary:	132
12	Appendix 4: Data Mining Implementation	135

Table of Figures:

Figure 1-1: Research Outline.....	7
Figure 2-1: (a) Probability density function of estimation ratios of all submitted bids for all projects. (b) Probability density function of estimation ratios of winning (lowest) bids for all projects (Chaovalitwongse et al., 2012).....	17
Figure 2-2: Model Process with Combination of Text and Numerical Data (Williams and Gong, 2014)	19
Figure 2-3: Data mining process when applied to software development processes. (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).....	24
Figure 2-4: Impact of variables and uncertainty based on project time (Schoonwinkel, Fourier and Conradie, 2016).....	29
Figure 2-5: Sustainable Project Management Circle:	30
Figure 2-6: Synthesised Data Mining Process	39
Figure 3-1: Data Mining Process and Chapter 3 Overview	41
Figure 3-2: Client-Engineer-Contractor Relationships.....	51
Figure 3-3: Data Pre-processing	55
Figure 3-4: Clustering and Outlier Detection Example.....	66
Figure 3-5: Decision Tree Donor Example (Aggarwal, 2015).....	70
Figure 3-6: Neural Network Example (Aggarwal, 2015)	72
Figure 3-7: Support Vector Machine Example (Han, Kamber and Pei, 2012).....	74
Figure 3-8: Linearly Inseparable Data SVM Example (Han, Kamber and Pei, 2012).....	74
Figure 3-9: Acetaminophen and its graph representation (Aggarwal, 2015).....	77
Figure 3-10: ROC curve for Outlier Analysis Validation.....	83
Figure 4-1: Visual representation of 9 clustering algorithms provided by Scikit-learn applied to 6 different unlabelled datasets (Pedregosa et al., 2011)	93
Figure 4-2: Scikit-learn algorithm cheat sheet (Pedregosa et al., 2011).....	94
Figure 4-3: Parsing sentence structure using NLTK (Loper and Bird, 2004).	95
Figure 4-4: Orange Visual Programming (Demšar et al., 2013).....	98

Figure 5-1: Final cost vs Employment Opportunities created	104
Figure 5-2: Tender cost estimate vs employment opportunities created	105
Figure 5-3: Tender cost estimate vs Employment Opportunities created (Cleaned data)	107
Figure 5-4: Cost (%) vs Number of Variation Orders	112
Figure 12-1: Implementation of Data Mining Application (Section 5.2)	139

List of Tables:

Table 2-1: Data Mining in the Construction Sector Case Study Summaries:.....	31
Table 2-2: Comparison of Data Mining Processes	35
Table 3-1: Classification assessment example.....	84
Table 3-2: Results of Cost Overrun Prediction Example (Lee et al., 2011):.....	89
Table 5-1: Employment Opportunities Creation Prediction Accuracies:	109
Table 11-1: Data Mining Resource Evaluation	134
Table 12-1: Sample of Objective Transformation	135
Table 12-2: Tender Cost Estimate Standardisation	136
Table 12-3: Employment Opportunities Binning:	137
Table 12-4: Target Dataset Sample.....	138

1 Introduction:

1.1 Introduction and Background:

“Data mining is the process of extracting knowledge from large volumes of data and selecting relevant information that is important for the decision-making process” (Syvajarvi and Stenvall, 2010). Data mining has seen a rapid increase in application to the communications, retailing, insurance and medical sectors. The uses range from fraud-detection to drug testing and customer retention. Data mining helps reduce costs, increase sales, and enhances research and development capabilities. These key competitive advantages allow data-driven organisations to deliver high quality, low cost, and short time-to-market products (Pospieszny, 2017). Organisations that have successfully applied data mining and have switched to data-driven decision-making processes have achieved, on average, 5% more productivity and 6% more profitability than their market competitors (Mcafee and Brynjolfsson, 2012). The application of data mining and data-driven decision making has been aided by the vast, and ever increasing, amounts of data stored by organisations about their internal processes, their products and their customers.

Project management is mainly concerned with delivering new products and services within an initially estimated budget and time frame. Despite the uncertainty throughout a project, the risk involved, and the need for accurate cost and time estimation data mining has not been widely applied to project management. The software development sector has been the first sector to adopt data mining to aid project management as the sector suffers from high project cost and time overruns. 53% of software development projects will cost 190% of their initial estimates and a third of all projects will be cancelled before completion (The Standish Group, 2014). These significant challenges and the complexity of applying traditional estimation techniques along with the large amount of data stored about each project has driven the adoption of data mining specifically for improved project management.

Construction projects also experience cost and time overruns. These overruns are a global problem but present a major challenge for developing countries. The highly competitive nature of the modern construction industry is placing increasingly complex demands on construction projects while still requiring the delivery of the project within the stipulated timeframe and cost. It is even more important that construction projects are delivered within budget in developing countries where the construction industry is a large economic driver and is focused on infrastructure and service delivery (Mukuka, Aigbavboa and Thwala, 2015; Senouci, Ismail and Eldin, 2016; Niazi and Painting, 2017). The importance of good construction project management is critical as the global construction industry is set to grow to 10.1 Trillion dollars in 2021 and the African and Middle-Eastern sectors outpacing the other sectors in terms of growth (IHS Economics, 2013).

During the lifetime of a construction project a large amount of information is generated. The information can be captured in formats such as project documentation, project databases and financial transaction information. A large amount of varied project documentation is generated. The project documentation serves many different purposes, from establishing

contracts between clients and engineers to informing clients of delays that have been experienced on site. A significant amount of work and attention goes into the creation of the project documentation and it represents all the official, and some unofficial, communication between the different parties involved in the construction project. Governments, engineering firms, large construction firms and private client bodies that are involved in many construction projects typically maintain a project database that captures information about the construction projects they have been a stakeholder in. The information stored in these databases will vary depending on the role of the organisation.

The application of data mining to the construction sector has been limited with data mining to aid project management being scarce. Several investigations into data mining in the construction sector have been conducted. These applications focused on specific goals, such as estimating the residual value of construction equipment or evaluating the best tender award policy (Fan, Abourizk and Kim, 2008; Chaovalitwongse *et al.*, 2012). Their varied goals and high success rates indicate that the construction industry has sufficient stores of data to be suitable for data mining.

If similar increases in productivity and profitability, as mentioned by McAfee and Brynjolfsson (2012), can be achieved in the construction industry by the application of a data mining, it would represent an enormous competitive advantage. The aim of this investigation is therefore to determine a process by which data mining can be applied to project management in the construction sector and to demonstrate the process on a real dataset.

1.2 Research Aims and Objectives:

Data mining is a.) the process of discovering patterns in large datasets involving methods at the intersection of machine learning, statistics, and database systems and b) the process of extracting knowledge from data for purposes of reporting relevant information to be used in decision making. (Syvajarvi and Stenvall, 2010).

The application of data mining to project management in the construction sector will require data scientists i.e. those familiar with the statistics, machine-learning and programming required to implement data mining algorithms. This requirement, together with the large variety of available applications, data sources, data mining algorithms and techniques, can be overwhelming for anyone without a background in the field, notably construction industry personnel.

The research has two main aims, each with its own objectives:

- **Aim 1:** To establish and describe a data mining process that can guide construction sector personnel in the application of data mining to construction project information to facilitate project management.
 - a. **Objective 1:** Synthesise a data mining process for application to project management in the construction sector.

- b. Objective 2:** Discuss the data mining process defined in **Objective 1**.
- **Aim 2:** To demonstrate and evaluate the defined data mining process by applying the process to a real project dataset obtained from the Directorate of Construction and Maintenance of the Western Cape Government's Department of Transport and Public Works.
 - a. Objective 3:** Discuss the available data mining resources necessary for the research and construction personnel to create a data mining application without the need to implement the machine-learning algorithms and data mining methods from scratch.
 - b. Objective 4:** Create and evaluate two data mining applications by following the data mining process defined and described in **Objective 1** and **Objective 2** by utilising the data mining resources described in **Objective 3**.

1.3 Scope and Limitations:

Data mining and construction project management are both broad fields with significant complexity. As such, the research introduces a scope with set limitations in this section to focus the investigation and to provide boundaries. The research aims to establish, describe and demonstrate a data mining process for application to construction project management for use by trained construction industry personnel. As such, the investigation assumes a familiarity with the construction sector and civil engineering but minimal familiarity with data mining. The scope and limitations are presented for each of the research objectives below:

Objective 1: The exact process used to implement a data mining application varies from application to application. The research will synthesise a broadly defined process to encompass all data mining applications in the construction sector towards improving project management.

Objective 2: Data mining contains a large number of algorithms and techniques that are used at different stages of developing a data mining application. The research will discuss the general data mining types, such as clustering and regression modelling. It will refer to the names of specific algorithms that performs these functions but will not provide in-depth explanations of these algorithms. The research will provide enough information to familiarise construction industry personnel with the main data mining concepts and possibilities. However, since there are numerous sources providing implementation details, these will not be given.

The project management requirements of a project are project specific and will depend on the goals of the project, the size of the project, the project management techniques employed and more. The research therefore addresses the mode of improving project management using data mining in a broad sense that encompasses general project management principles. These may then be expanded according to the specific application.

Objective 3: The research will provide a number of possible resources that can be used to implement data mining. The resources will be mentioned, and their main capabilities discussed. The purpose of this is to inform construction industry personnel about the available resources and to enable them to decide which resource to base their application on, depending on their needs and skillset. The resources all contain detailed implementation guides, tutorials, and explanations of the algorithms they provide. Therefore, the research will not repeat these details and will instead refer the reader to the resource (generally a website or online store).

Objective 4: The demonstration of the data mining process applied to the data set provided by the Western Cape Government's Department of Transportation and Public Works will be conducted with selected resources from **Objective 3** and will not demonstrate all the resources. The reason for this is that several resources have the same capabilities but are implemented in different programming languages or are paid-for applications. The application here will follow the data mining process defined and discussed in **Objective 2** and **3**. The rationale for deciding which type of data mining use will be discussed, along with the exact algorithms used. The data mining application will not be an implementation guide. The outcome of the data mining application alone should not determine whether the data mining process was useful since a good process can nevertheless lead to unsuccessful applications. Unsuccessful applications could be the result of an over-ambitious goal, insufficient data or environment complexity. The completeness of the process synthesised in **Objective 1** will be discussed and changes, if necessary, suggested.

1.4 Methodology:

This section introduces and briefly describes the methodology the research adopted to achieve the aims and objective of the study within the defined scope and limitations. The methodology used to achieve each objective is discussed below:

Objective 1: Synthesis of a Data Mining Process for use in the Construction Sector to Improve Project Management:

Qualitative research processes such as a literature review and the examination of case studies were used to synthesise the data mining process. Case studies of data mining in the construction sector and in the software development sector were obtained from accredited journals, research reports and publications. Case studies are used because data mining in the construction sector is relatively rare and no suitable candidates could be identified for surveys or interviews. Eight case studies in total were chosen, based on their contribution of new knowledge to the data mining process and how that might be applied to facilitate project management. The case studies of data mining in the construction sector could not provide all the required information, specifically about how to improve project management using data mining. This was the primary reason for considering data mining case studies from the software development sector.

The case studies were examined for information regarding the specific data mining process they utilised, along with practical implementation information. Other information that could influence the application of data mining to the construction sector was extracted from several accredited journals, books, publications and internet sources and are discussed.

The information gained from the case studies was combined with the knowledge gained from the surrounding literature to deductively synthesise a data mining process that is specifically aimed at improving project management by mining construction project data.

Objective 2: Description and Discussion of the Data Mining Process:

The data mining process synthesised in **Objective 1** is described and discussed in **Objective 2**. The technical aspects of the data mining process are described using information summarised from two leading data mining textbooks. These textbooks were chosen as they form the basis of many post-graduate courses in data mining and data science, such as at Hong Kong University of Science and Technology and at the University of Illinois and have been cited in many accredited journals.

The discussion of non-technical aspects of the data mining process draws on information presented in the literature review and is supplemented with information from journals, textbooks and dissertations.

The description and discussion of the data mining process is conducted without discussing in-depth technical and implementation detail in order to provide construction personnel or indeed, any other interested party, with basic knowledge of the field.

Objective 3: Presentation and Discussion of Data Mining Implementation Resources:

The implementation of a data mining application is typically done via some form of computer programming and analysis. The implementation of data mining methods from first principles can be extremely complex. For this reason, the research presents and discusses some data mining and machine learning resources that lower the barrier of entry for a novice data mining practitioner. However, most of the available resources require the user to have some basic programming knowledge.

The resources presented are widely used and have been developed by both professional and amateur data mining and machine learning practitioners. The selection of resources was based on their user-friendly nature without sacrificing accuracy or mathematical rigor. The resource costs are considered along with any associated end-user licence agreement that might prohibit application in a commercial context.

Objective 4: Demonstration of the Data Mining Process on a Real Dataset:

The final objective of the research is to demonstrate the data mining process on a real project dataset. This is achieved by applying the data mining process to a dataset of 755 projects. The dataset was obtained from the internal project database at the Road Construction and Maintenance Directorate of the Western Cape Department of Public Works in South Africa.

This step is mainly qualitative in nature, despite the use of quantitative techniques. This is due to the fact that the results of the data mining application are not directly used to determine whether the process is valid. Instead, the validity of the process will be quantitatively examined by discussing the application and any omitted or additional information required.

1.5 Research Outline:

The outline of the document, beyond the introductory Chapter 1, is presented in Figure 1-1 with a brief discussion of the content of each chapter.

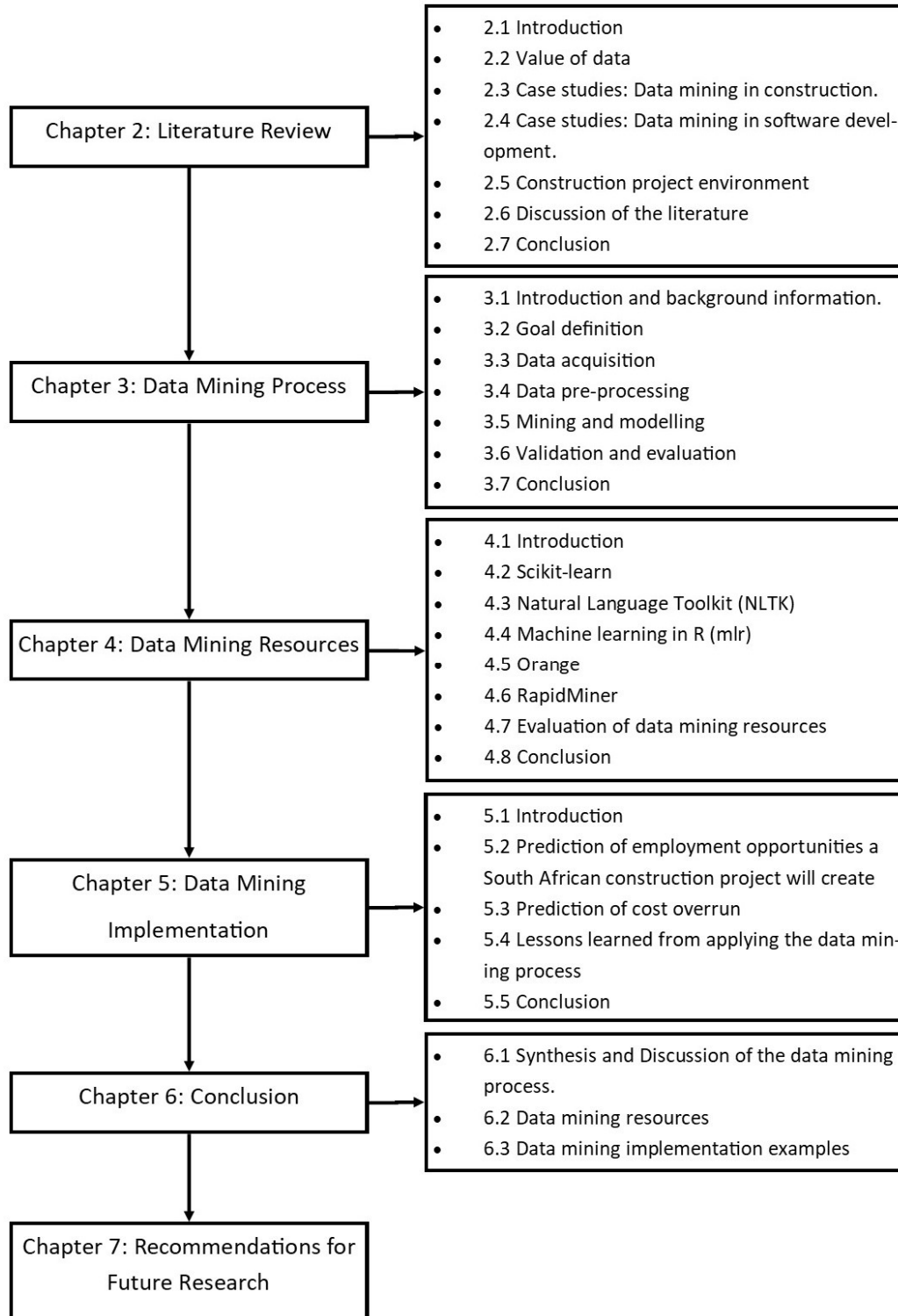


Figure 1-1: Research Outline.

1.5.1 Chapter 2: Literature Review:

The literature review investigates eight case studies of data mining in the construction sector and data mining applied to software development projects. The value of accurate data and estimates are discussed along with construction project phases and the sustainable success criteria used within the construction sector. The suitability of using data mining in construction projects for improving construction projects is discussed. Finally, a data mining process is synthesised for use in facilitating the management of construction projects.

1.5.2 Chapter 3: Data Mining Process:

The data mining process synthesised in Chapter 2 is discussed in detail in this chapter. The goal of each step within the data mining process is discussed and methods for achieving these goals are presented. Where possible, the discussion focuses on the application of data mining to construction projects to familiarise construction sector personnel with the data mining process and its many possibilities.

1.5.3 Chapter 4: Data Mining Resources:

Data mining resources that focus on the user-friendly, yet rigorous, implementation of data mining is presented in this chapter. Five resources are presented that cover implementation of data mining in two programming languages and both free and paid-for software packages. These resources are presented to enable relative data mining novices to develop a data mining application.

1.5.4 Chapter 5: Data Mining Implementation Example:

The data mining process is applied to a real project dataset obtained from the Western Cape Government's Department of Transportation and Public Works. A data mining application is developed using the data mining process and the results of the application are discussed. A number of lessons learned during the implementation are discussed to provide useful extra information for data mining novices.

1.5.5 Chapter 6: Conclusion:

A summary of the research and conclusions reached therein are presented.

1.5.6 Chapter 7: Recommendations for Further Research:

A number of opportunities for further research are identified into data mining in the construction sector. These include possible large-scale applications to demonstrate the value of the technology to construction, the training required for construction sector personnel to adopt data mining and the possibility of building information models as data sources.

2 Literature Review:

2.1 Introduction:

Data mining is the process by which patterns are discovered in large datasets by the application of techniques that span the fields of machine learning, statistics, and database systems. The knowledge extracted is reported for the purpose of facilitating data-driven decisions (Syvajarvi and Stenvall, 2010).

In this chapter the literature is examined in order to define a data mining process that can be used in applications designed to aid the management of construction projects.

The value of data is examined within this chapter to investigate the relationship between having accurate data, data-driven decision making, and the real-world gains which organisations have achieved by ensuring their decisions are based on accurate data. The impact of poor-quality data is discussed on the operational, tactical and strategy level of an organisation.

The literature was consulted for examples of data mining applied to construction projects towards better project management. These examples prove to be scarce, leading to a broader goal of data mining throughout the construction sector being examined. Several case studies were found that provided useful information into possible applications of data mining including the breadth of methods and techniques and the variety of data that can be used.

The investigation then turned to the software development sector in an attempt to obtain information that focused specifically on data mining for improved project management. Software development projects encounter many of the same problems that construction projects face (cost overruns, time overruns etc.). Since data mining has been applied successfully to the software development sector, it provided valuable information to this investigation.

Construction projects were examined to ascertain if the similarities between them and software development projects are sufficiently significant to warrant the application of data mining. Both the project phases and the levels of uncertainty within each phase were compared. The core success criteria of construction projects were also examined to determine their contribution to the uncertainty within a construction project. For example, the criteria of cost, time, and quality produce considerable uncertainty within any project.

A summary of the knowledge obtained and a discussion thereof is presented. The knowledge obtained about construction projects, data mining applied to the construction sector, and data mining applied to software development projects with the specific aim of facilitating project management were used to synthesise a complete data mining process. Although the data mining process defined here was specifically designed for data mining applied to construction projects, with slight modification it may be applied for the purpose of facilitating project management in any sector.

2.2 The Value of Data:

“That which does not get measured, does not get managed” (Redman, 1998). Data is a valuable source of knowledge for making project and business decisions. Provided that the data is both accurate and representative, it can be a critical basis for operational, tactical, and strategic levels of decision-making. The value of having good data and what that means to data driven companies is discussed in this section. The real-world gains of data driven companies are also discussed along with the impact of using poor or insufficient data.

2.2.1 Value of Data and Data Driven Companies:

Data is the basis of knowledge. Having more data about a process, project or a company allows one to more accurately understand and model it. This increased accuracy can yield real-world gains in efficiency, performance and more accurate planning.

Data driven companies are those that base their strategic, tactical, and operational decision-making on accurate and abundant data. Companies that have implemented large scale data collection, analysis and data-driven decision making processes are, on average, 5% more productive and 6% more profitable than their market competitors (Mcafee and Brynjolfsson, 2012).

Many large retail stores are already using ‘Big Data’ techniques to track and analyse as many different data points about their customer base as possible. The data ranges from items bought and shopping times to age and gender. Using this information and very specific data-querying and modelling methods they are able to predict what types of items would interest a prospective customer. This information is then used to target their advertising campaigns.

The airline industry uses data from pilots, past flights, current weather reporting and other sources to predict, with very high accuracy, the arrival times of planes. This allows airports to schedule planes landing within minutes of each other, thus increasing efficiency and revenue (Mcafee and Brynjolfsson, 2012).

2.2.2 Impact of Poor Data:

Poor quality data has a negative impact on customer satisfaction; effective decision-making; increases operational cost and reduces the ability of an organisation to formulate and execute strategy. Some of the less quantifiable impacts include lower morale, mistrust of the organisation, and difficulties aligning the organisation (Redman, 1998). The wide variety of data quality issues fall into one of four broad categories:

- Models of the real world captured in data (data views). These include issues with granularity, relevancy, and level of detail.
- Data values. These include issues of accuracy, completeness, and consistency.
- Presentation and reporting of the data. These issues include the ease of interpretation, the suitability of presentation format and loss of detail.
- Privacy, security, and ownership of the data.

To reiterate Redman (1998), “That which doesn’t get measured, doesn’t get managed.” Data which is not captured or is inaccurate, which is typically between 1% and 5% of all data for organisations that do not have any special data quality checks, can have negative impacts on an organisation on an operational level, tactical level and strategic level.

On the operational level, poor data impacts on customer satisfaction, increased costs and lowered employee job satisfaction. Customers will become dissatisfied if their bills, orders or deliveries are incorrect or late. Many customers expect these details to be correct and are very unforgiving of mistakes. It costs money and time to fix errors made due to poor data quality. Wrong orders and deliveries can increase operating costs as extra work has to be done to rectify the issue. Resources must also be spent on detecting and rectifying issues in data. Employee job satisfaction is lowered when they are placed under pressure to fix issues or are forced to deal with dissatisfied customers. Studies to estimate the total cost of poor data have been difficult to perform but three proprietary studies have produced a figure of between 8% and 12% of revenue lost due to poor data (Redman, 1998).

Impacts on the tactical level are just as significant as the operational level, although they do not carry the same monetary value as the impacts of poor data quality on the operational level. There is no significant evidence that the data used by managers is of a higher quality than the data used by customer service employees. Poor quality data will therefore influence decision making as decisions are typically only as good as the data they are based on. While some uncertainty is present in all decisions, it is clear that better quality, more accurate, more relevant and timely data will lead to better decisions. Poor data quality can also lead to mistrust within an organisation as each department has its own data that may be inconsistent with another department, increasing the difficulty of cooperation within the organisation (Redman, 1998; Borek *et al.*, 2013).

The selection, development and evolution of an organisation’s strategy is itself a long and continuous decision-making process and thus the impacts that can be seen on lower levels of the organisation carry up to the strategic level. The lack of accurate data on the market, customers, competitors, new technologies and other salient factors of the environment in which the company operates makes it difficult to formulate a sound corporate strategy. Corporate strategy dictates the short-term, medium-term and long-term plans. As these plans are rolled out they are assessed and modified based on results obtained. If the reported results are inaccurate or unreliable, it can dramatically affect the execution of the corporate strategy (Redman, 1998; Borek *et al.*, 2013).

Construction and the built environment is not exempt from the same types of data problems that other sectors experience. However, similar advantages could be afforded to organisations within the construction sector if they embrace the data revolution and ensure that their decisions are based on good, accurate data.

2.3 Case Studies of Data Mining and Text Mining Used in Construction and Project Management:

This section looks at case studies of data mining applied to the construction sector. The available literature was examined to determine what investigations have been conducted into data mining in the construction sector.

Five case studies were examined to extract the following information:

- Possible data mining applications.
- The requirements for applying data mining to the construction sector.
- Possible data sources.
- Other useful information about regarding a general data mining process for facilitating construction project management.

Furthermore, since the case studies each applied different data mining models to varied applications using unique processes, they are later summarised and compared to determine if construction projects are suitable candidates for data mining. The summaries are used in the synthesis of the data mining process for improved project management.

While each of the case studies are examined in detail, the exact model types and methods are deemed as important as the overall process, the types of data used, the data source, and the goals of the overall application. Of the case studies presented below, there are no specific examples of data mining used to improve construction project management. While the information gained from some of the case studies could prove useful for project management, this was not the specific goal of these applications. Either there have been no such applications, or none have been published. Hence, the examination of the software development sector for case studies that specifically aim to help project management through data mining of project information in Section 2.4.

2.3.1 Case Study 1: Residual Value Assessment of Heavy Construction Equipment:

Fan, Abourizk and Kim (2008) used predictive data mining and a national database of construction equipment in order to assess the residual value of heavy construction equipment. The total value of the construction equipment in the USA, at the time of their publication, was over US\$ 100 Billion. In order to minimise the equipment cost per unit of service, a contractor must make important decisions about equipment acquisition, replacement, repair, and disposal on a regular basis. The residual value, or current market value, of the construction equipment is cited as one of the most important factors when making those decisions (Fan, Abourizk and Kim, 2008). Since the current market value of equipment can only really be assessed when the equipment is sold at auction, making such an estimate is difficult.

The data source used by Fan, Abourizk and Kim (2008) was Last Bid, a US based online construction equipment database covering up-to-date auction results across the US and other international markets. Information was gathered from auctions for heavy construction equipment held between 1996 - 2005 that included the make, model, model year, auction year, condition, location, and auction price. Other information that could influence the price of heavy equipment, such as yearly construction investment and gross domestic product was gathered from Statistics Canada and the US Bureau of Economic Analysis.

The residual value of heavy construction equipment is influenced by various features. To ensure model accuracy, all the features that can significantly influence the outcome must be selected and added to the model. Some features were transformed, either to improve the accuracy of the model or to fit the input format of the model. Two examples of transformed features are the equipment age and the auction location. The equipment age was calculated as the difference between the year of make and the auction year. This translates to the working life of the equipment at the date it was sold. In the auction database, the auction location was recorded as region, state, or county. The authors standardised the location to refer to only the region.

The quality of the input data significantly influences the quality of the knowledge generated from the data. Consistent formatting, removal of outliers, and removal or filling of missing values is vital to generating an accurate model. In addition to these general requirements, it is important that the data should be representative of the full range of all the features that appear in the data set. Under-representation of certain features will result in poor accuracy for that feature. In this study, an example is the absence of an auction region or a specified price range. Entries with missing values can have values assigned to the entry or the entry may be deleted, according to which specific data is missing. In order to reduce the complexity of the data mining algorithm, continuous variables, such as price were binned into discrete variables (Fan, Abourizk and Kim, 2008).

Data mining to predict a numerical value is a common data mining task, where the most likely value of a response variable is determined based on the known predictor variables, or features. The generalised form can be given as: $y = f(x_1; x_2; x_3; \dots; x_n; r_1; r_2; r_3; \dots; r_n)$. Where y is the continuous target variable, x_i ($i = 1, 2, 3, \dots, n$) is a predictor variable and can be either categorical or continuous and r_i ($i = 1, 2, 3, \dots, n$) is a model parameter. The $f()$ represents a data mining model's discovered patterns or rules, which are learned from data inserted during the 'training' period. While it is theoretically possible to build a single model that is capable of predicting the residual value of all types of heavy construction equipment, provided that all the varieties are adequately represented in the training data, a model of that scale would both have poor quality and be difficult to interpret (Fan, Abourizk and Kim, 2008).

The authors therefore decided to create several models, one for each category of heavy construction equipment. The authors decided to use an Auto Regressive Tree Algorithm (see Section 3.5.5 for an overview of Decision Trees and Regression) as it establishes non-linear relationships between variables. Whereas this particular model allows the user to examine the

relationships within the data, based on the patterns it has learned, this is not always possible as other data mining models may have unintelligible classification methods.

The database for each model was split into training and testing datasets. The model was trained on the training datasets and then evaluated for accuracy and reliability using the testing dataset. Machine learning algorithms that learn the internal data patterns based on a ‘training’ dataset and are then required to predict values for a ‘testing’ dataset are known as supervised machine learning algorithms (see Section 3.1.2).

A ‘ten-fold’ cross validation (see Section 3.6.3.3) was conducted to validate the results of the classifier and to ensure they are accurate. After the model was validated, the usability of the model was tested by developing a real time price prediction model that could be accessed via a website. By providing information about a specific piece of heavy equipment, a website visitor would receive a prediction of its residual value. The model made the prediction fast enough to be usable by a customer or a client. The rapidity with which the model trains allowed the authors to set it up to retrain nightly with data from the auctions of that day added to the dataset. As more data is collected, the model’s accuracy increased over time (Fan, Abourizk and Kim, 2008).

The investigation by Fan, Abourizk and Kim (2008) is promising for the purpose of using data mining and predictive models in the construction sector. Further, the authors demonstrated the necessity of re-examining the goals of the application for possible adjustment once the data has been collected. They stress the importance of carefully selecting, transforming and preparing the data; rigorously testing the accuracy of the model prior to deployment and ensuring that the application is practical and usable.

2.3.2 Case Study 2: Using Data Mining to Discover Knowledge in Enterprise Performance Records:

“Data mining is one of the core methodologies of knowledge discovery in databases” (Lee, Hsueh and Tseng, 2008). Lee, Hsueh and Tseng set out to demonstrate a data mining application in the construction industry where, rather than predicting values, the goal was to acquire knowledge about a process or activity.

Data mining is both able to automatically analyse information in databases and attempt to interpret the information into new knowledge (Lee, Hsueh and Tseng, 2008). By applying recursive iterations, a data mining algorithm can classify the data into predefined groups. The model learns from examples and uses characteristics of the data to classify the data. The authors used a Decision Tree Classification algorithm (see Section 3.5.5) for its capacity to manage both discrete and continuous information; generate and demonstrate comprehensible rules; and identify the level of significance of independent variables. This allowed the authors to determine causes of poor quality in building construction from the information in the construction databases.

After-construction maintenance data from 1994 - 1997 was collected from the service and maintenance department of a large construction company in Taiwan. This amounted to 7790

cases divided into 35 service categories. Since there was ample data about the two main causes of call backs i.e. leakage and cracking, the authors decided to focus on those factors in an effort to reduce maintenance costs for the company. The data collected from the maintenance records was both disjointed and stored across several databases. In order to increase the possibility of discovering underlying causes of leakage or cracking that may have arisen in the design or construction phases of the project, the authors compiled a target dataset by combining information from these phases with the information from the maintenance database.

The following data mining process was applied:

- Step 1: Data selection.
- Step 2: Data cleaning and preparation.
- Step 3: Data reduction and coding.
- Step 4: Algorithm selection.
- Step 5: Mining and reporting.

The process flow presented above is linear, but the authors conducted numerous iterations, adjusting their application to increase its accuracy and efficiency. Key business data, such as payment information, was meticulously recorded in the databases but data about project scope and execution details has not been as rigorously recorded during the execution of the project. In a laborious bid to improve their target dataset the authors added design and construction phase information to their dataset. Since the patient and careful preparation of the target dataset will significantly influence the accuracy and dependability of the information extracted from the data, the authors argue that this is the most important factor that an investigator can control (Lee, Hsueh and Tseng, 2008).

The refined target dataset was mined using the selected Decision Tree Classification algorithm and the authors were able to extract valuable information about cracking and leakage. Three main recurring factors in cases of leakage and cracking were identified: high-strength concrete, high rise construction (especially above the twentieth floor), and the name of a specific site manager. The extracted knowledge was examined by the company and its engineers before reaching the following conclusions:

- High-strength concrete segregated when pumped to heights of twenty stories and above due to the use of high powered pumps. The segregated concrete resulted in a porous structure prone to leaking.
- To facilitate pumping to heights in excess of twenty stories, the concrete was mixed with a high-water content, resulting in increased shrinkage leading to cracking and leaking.
- More severe wind conditions and direct sun exposure from the twentieth floor and above increased the water evaporation rate when the formwork was removed leading to concrete shrinking and cracking.
- Insufficient training of on-site personnel led to poor quality work prone to cracking and leaking.

The results of a data mining application are dependent on the quality and variety of the input data. Only those causes of cracking and leakage already present in the data could possibly be identified. The authors argue that this in no way diminishes the usefulness of data mining. Rather, it allows individual enterprises to examine their records to discover the specific reasons for their failures (Lee, Hsueh and Tseng, 2008).

This investigation shows that data mining can extract valuable knowledge from existing project information, even about a process that is highly understood and often repeated. The authors followed a structured data mining process that they repeated to improve their results. The authors stress that proper attention to setting up a target dataset is one of the most important factors in data mining. In this case study, data was gathered from various databases to form a comprehensive target dataset. The authors also show that the results from data mining might require human interpretation and may not necessarily be immediately usable, as was the case in Case Study 1.

2.3.3 Case Study 3: Data Mining Tender Bid Information to Evaluate the Best Bid Selection Policy:

Cost overruns on construction projects are a common problem for the industry. Among the factors influencing cost overruns is the extremely competitive construction marketplace (Chaovalitwongse *et al.*, 2012). By law, many governments and other agencies are compelled to use the Lowest Bidder tender policy for construction projects. Non-governmental organisations are not required to utilise the same tender policy, but many still select the lowest bidder as their preferred bidder. As a result, some construction companies submit bids that are lower than the actual cost of the project and rely on claims, change orders and other disputes to make the project profitable. The net result is that project owners incur significant cost overruns (Chaovalitwongse *et al.*, 2012).

There are several other bid selection policies, such as the Second-Lowest and Trimmed Mean tender selection policies. These selection policies aim to reduce the possibility of selecting bids that are not close to the actual cost of the project. This paper aimed to use data mining and machine learning to examine the bid policies and create a machine-learning application that selects the bid closest to the actual price of the project.

The bid data used in the paper was obtained from the Texas and California Departments of Transport, which contained approximately 4000 projects in total. The data contained construction information along with information about the bids for each project. Data from projects where there were extremely large cost overruns or underruns was removed from the dataset. The authors' reasoning for this was that those projects most likely had some large increase or decrease in scope. The authors conducted a statistical analysis of the data prior to applying the data-mining algorithms to the dataset to determine the variation in bid estimates and to ensure that the data collected was usable and not heavily skewed. The distribution of the estimation ratios for the submitted bids and the selected bids are shown in Figure 2-1, where the estimation ratio is the submitted bid amount minus the actual amount divided by the actual amount.

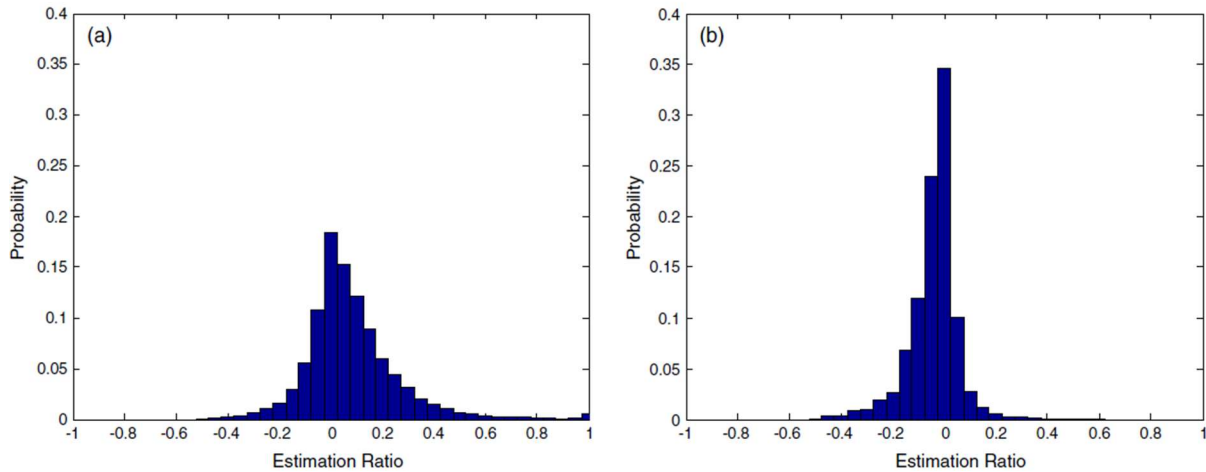


Figure 2-1: (a) Probability density function of estimation ratios of all submitted bids for all projects. (b) Probability density function of estimation ratios of winning (lowest) bids for all projects (Chaovalitwongse et al., 2012)

The authors determined that for all bids received, 54% were within 10% of the actual cost of the project, whereas 77% of the selected bids were within 10% of the actual cost. The authors applied 5 bidding ratios to describe the patterns within the data for each project:

- Mean-Bid Ratio: $= \frac{\text{Mean Bid} - \text{Low Bid}}{\text{Low Bid}}$;
- Median-Bid Ratio: $= \frac{\text{Media Bid} - \text{Low Bid}}{\text{Low Bid}}$;
- Maximum-Bid Ratio: $= \frac{\text{Maximum Bid} - \text{Low Bid}}{\text{Low Bid}}$;
- Coefficient of Variance: $= \frac{\sigma}{\bar{x}}$;

The decision to use the data for the investigation was based on the large quantity available and the fact that it followed a typical Gaussian distribution. The only change made to the investigation involved splitting the data into small and large projects to offset the influence of the distributional differences that existed between the two groups. The cut-off point of \$100 000 for small projects was chosen to isolate the two distributions and to split the dataset into two roughly equal subsets.

Two types of Neural Networks (a group of machine-learning algorithms discussed in Section 3.5.5) were applied to the dataset. The algorithms selected were a Probabilistic Neural Network (PNN), which allows for Neural Network Classification and Neural Network Regression modelling, and a Generalised Regression Neural Network (GRNN), which allows for Neural Network Regression modelling (see Section 3.5.5). The problem was modelled as a classification problem (the algorithms had to select an optimal bid). The PNN was used in its classification modelling set-up whereas the GRNN provided an optimal bid amount and selected the closest bid. The 5 bid ratios for each project was the input data into the model. The neural networks were trained and validated by repeating a 5-times cross validation process 10 times to ensure valid results.

The Neural Network selected bids were compared with the bids selected by lowest bid, second lowest bid, mean bid and trimmed mean bid policies. The evaluation of the different bid policies and the Neural Networks is discussed in detail by the authors, taking into consideration the construction environment and its intricacies. The counter-active influences of clients wanting to pay the least for a project and not wanting the project value to overrun the estimated value adds layers of difficulties when setting up bid selection policies.

The authors conclude that the traditional policies that were the most effective at balancing these influences are the lowest and second-lowest bid selection policies. Of the Neural Networks models, the PNN achieved the best success by matching these two traditional bid selection policies. The authors argue that with more training data and with finer tuning, the PNN will outperform traditional bid selection policies (Chaovalitwongse *et al.*, 2012).

This investigation shows that data mining can be used to evaluate current policies and is not limited to predictions or knowledge extraction. The authors conduct a preliminary data examination to determine if the data will be usable for the investigation. The input data for the data mining algorithm is not the actual project bid data but rather ratios and information about the data. This is a critical insight as normalising data or utilising internal data ratios are often required to reduce overweighting of certain features.

2.3.4 Case Study 4: Predicting Construction Cost Overruns Using Data Mining:

Williams and Gong (2014) set out to predict construction project cost overruns using text mining and numerical data. A large variety of factors influence construction project cost overruns. Most previous cost overrun modelling attempts were made using only numerical data from the project. With the recent success and advances in text mining, the authors decided to use an approach which combined numerical and text data into one dataset. This case study uses several advanced data mining techniques that are all discussed in Section 3.5. The process followed by the authors is presented below to illustrate how complex data mining application can become, and the value of combining textual and numerical data rather than explain the process itself.

The data used in the study was collected from the California Department of Transport and contained numerical bid information along with a short paragraph describing the project's major work and cost items. The projects with extreme cost overruns or underruns were purged from the dataset, as these projects usually contained large scope changes. The projects were divided into three groups of cost overrun: projects with high cost overruns ($x > 6\%$), projects with medium cost overruns ($6\% > x > 3\%$), and projects with slight overruns or underruns ($x < 3\%$).

Several models were trained on 60% of the data and then tested on the remaining 40%. As shown in Figure 2-2, the process starts with splitting the data into the text and numerical parts which were processed separately. The text was transformed into a numerical representation (a numerical matrix) through a process of dividing up the text into individual words

(‘tokenising’) and reducing the words to their root form (‘stopping’, ‘stemming’, and ‘normalisation’). The text-processing output is an extremely large matrix, that is mostly empty, where all the unique words are represented by the columns and the projects are represented by the rows and where the value is the number of times each word occurs in each project (see Section 3.5.6: Text Mining). The numerical word matrix is collapsed into a numerical word vector by using Single Value Decomposition (SVD). The cleaned numerical values were combined with the numerical word vector into a target dataset. Four different classification models were trained and tested on the training and testing target datasets.

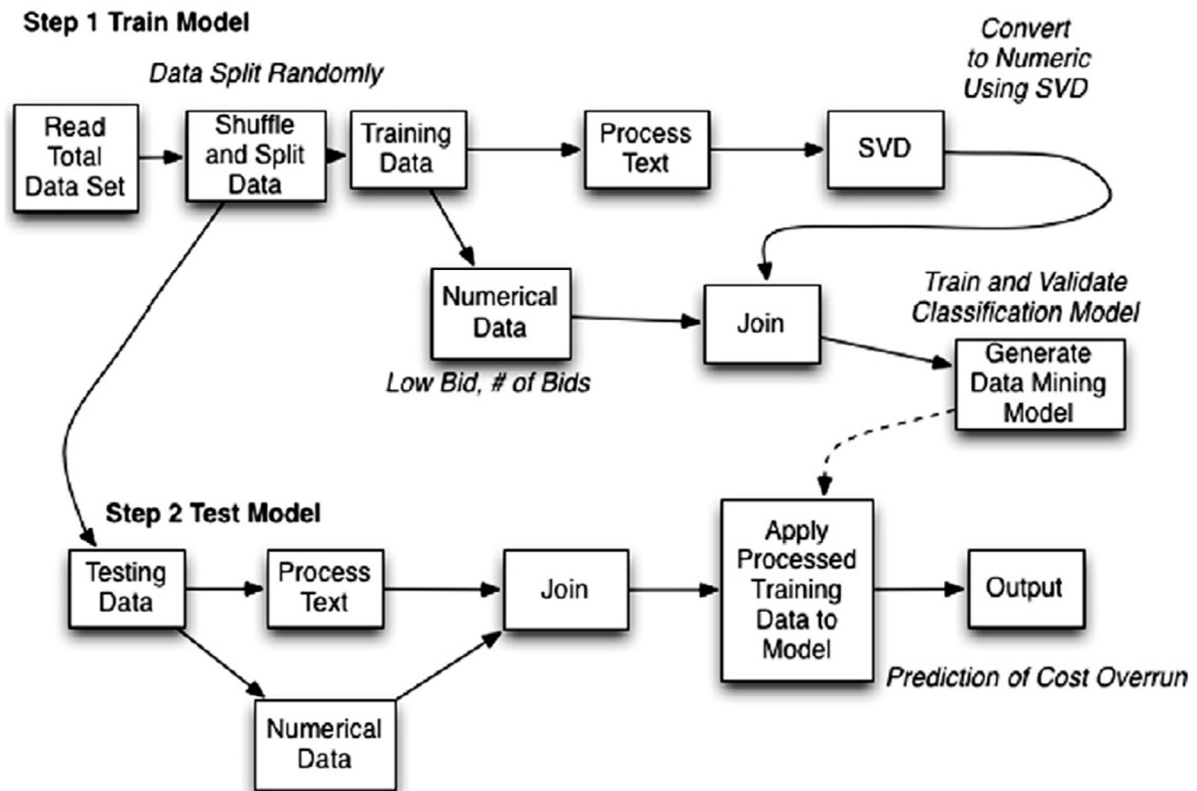


Figure 2-2: Model Process with Combination of Text and Numerical Data (Williams and Gong, 2014)

A technique known as bootstrapping, used when a limited amount of training data exists, was employed to increase the accuracy of the prediction. This entire training, bootstrapping and testing process was repeated five times to validate the prediction accuracy.

After the analysis process it is possible to determine which words are most associated with high cost overruns. Words such as “replac_bridg” and “excavat_ashphalt” were highly correlated with cost overruns. Using such words enables the authors to identify projects that run the risk of time delays or cost overruns. The overall prediction accuracy of the different models varied from 40% to 44%. This result was significantly lowered by the very poor accuracy of predicting projects with cost underruns. The classifiers were best able to identify projects with high cost overruns. When the same data mining models were trained on numerical data alone, they all produced poorer results in almost every prediction category.

This investigation showed that using numerical and text data together in the target database allows a more complete model of the projects to be constructed by the data mining algorithm, leading to more accurate predictions. The poor accuracy in predicting projects with low overruns and cost underruns points to the possibility that the construction environment may be too complex for accurately modelling in its entirety or that the study did not have sufficient data.

2.3.5 Case Study 5: Data Mining to Detect Early Warning Signs of Project Failure by Mining Unstructured Text from Site Meetings:

The complexity of the construction sector makes construction projects prone to failure from a wide variety of causes. The failures must be dealt with and the resulting time delay and cost increase negatively influence the project owners, the contractors and the whole project team. Project management methods are designed to help anticipate and minimise the project risks. However, these methods alone cannot guarantee project success.

Alsubaey, Asasi and Makatsoris (2015) set out to create an early warning system by mining unstructured text from project site meetings to identify signs of project risks materialising. Such a system would allow the project team to react more quickly when detecting possible risks materialising, to quickly rectify the issue and prevent possible time delays and/or cost increases.

The authors identified 10 categories of risk that an early warning system should be able to recognise. These are defined by a lack of: (Alsubaey, Asadi and Makatsoris, 2015).

1. Onsite materials.
2. Manpower
3. Keen commitment to the project milestones and scopes.
4. Stable project requirements.
5. Overall safety.
6. Making purchases.
7. Understanding of a new project.
8. Project team required knowledge/skills.
9. Due diligence on vendor(s) and team members.
10. Top management support or commitment to the project.

The authors acquired the site meeting minutes from 46 projects that had experienced delays or cost increases. The authors manually labelled a training dataset according to the ten categories they identified. A Naïve-Bayes Classifier (discussed in Section 3.5.5.3) was trained on the manually labelled training dataset and tested on the data from 46 unclassified projects. The early warning sign most commonly identified in the text was ‘lack of onsite materials’ with it being present in 80% of the test data. The second most common classification was ‘lack of keen commitment to project milestones and scopes’. These two categories made up the clear majority of the identified issues for the test projects.

By testing a project's site meeting minutes throughout the construction period, the project team will be in a better position to identify issues such as a lack of onsite materials and rectify them (Alsubaey, Asadi and Makatsoris, 2015).

This case study shows the value of using unstructured text as a data source for data mining. By applying text mining, valuable information regarding on-site issues that personnel might not have noticed were identified. This application shows the potential of applying data mining to a smaller issue within the construction sector provided that the application goal is cognisant of the data volume issues.

2.4 Case Studies of Data Mining Used to Improve Project Management in Software Development:

This section examines case studies of data mining in the software development sector to improve project management. Since no case studies of data mining aim at specifically improving construction project management were available, the research expanded its view to encompass case studies from other sectors where improving project management by data mining was the main focus. These were found in the software development sector.

The software engineering and development sector faces many of the same challenges as the construction sector in terms of project management, often to a greater degree. More than half of all software development projects will cost almost double their initial estimates and a full third of software development projects will be cancelled before completion (The Standish Group, 2014). These project difficulties and the sector's unique proximity to computer science specialists have resulted in several published examples of adopting data mining to assist in project management.

Three case studies are examined in this section for knowledge on the data mining process used and how to apply data mining to improve project management. The types of information that should be captured and the project phases during which data mining should be applied to benefit project management have been extracted from these case studies. The knowledge drawn from these case studies was ultimately combined with the knowledge acquired from case studies of data mining in construction as well as general information about construction projects to determine whether construction projects are suitable for data mining as well as to synthesise a data mining process. Several case studies were found but only three are examined here as the information presented in these three cases studies were repeated in the other case studies.

2.4.1 Case Study 6: Data Mining Application in a Software Project Management Process:

In the software development environment the two major difficulties facing a project manager are a.) accurately estimating the duration of work packages and the project as a whole and b.) estimating and managing problems and bugs that arise during the development process

(Nayak and Qiu, 2005). Nayak and Qui (2005) set out a comprehensive data mining process that they defined as follows:

1. Data Acquisition.
2. Data Pre-processing.
 - a. Defining Goals.
 - b. Field Selection.
 - c. Data Cleaning.
 - d. Data Transformation.
3. Data Modelling and Mining.
4. Assimilation and Analysis of Outputs.
 - a. Classification and Association Rule Mining.
 - b. Text Mining.
 - c. Addressing Errors and Problems.

Nayak and Qui (2005) followed this data mining process to implement their data mining application.

Data Acquisition: Their data was acquired from MASC, a division of a global telecommunication company, which recorded and stored 40 000 software problem reports from hundreds of thousands of lines of code. The problem reports contained information such as synopsis, severity, priority state, arrival-date, closure-date, description etc.

Data pre-processing: Defining the goals for the data mining is extremely important as it guides and informs the collection and preparation of the data, the data mining algorithms chosen, and the format in which the knowledge gained will be used or presented. Nayak and Qui (2005) defined their goal to be mining the data contained in the software problem reports to help the project managers make better time estimates for bug fixing.

The authors selected the information they thought would contribute to achieving their stated goal and included it in their target dataset. The data was cleaned of missing or erroneous values and some data fields were transformed into a more usable format, such as transforming arrival date and closure data into duration. Due to problems or inconsistencies in the data, only 11 000 problem reports from an original 40 000 were eventually used.

Data modelling and mining: The authors adopted a Decision Tree algorithm (discussed in Section 3.5.5.1) for part of their investigation. Decision Trees fall under the category of supervised learning, where the algorithm is given a set of labelled training data. This data, along with the desired output, is provided as examples from which the algorithm learns. The algorithm then uses a specific technique, in this case a Decision Tree, to model the input data with the input labels.

Once the model has been trained using the labelled training data, the model is tested on unlabelled testing data during which the model must assign a label to the testing data. The label given to the testing data is then compared with the known label of that data to check its accuracy. The training and testing datasets are created by splitting the pre-processed target

dataset. The authors also used text mining on the unstructured text data contained in the ‘synopsis’ and ‘description’ fields of the target dataset. Natural language processing techniques were used to pre-process the textual data in much the same way as was done in Case Study 4 by Williams and Gong (2004).

The suitability of the data-mining tools and methods used are judged on the ease of use, cost of use, ease of preparation and appropriateness for the data and goal of the data mining application (Nayak and Qiu, 2005).

Assimilation and analysis of outputs: The Decision Tree algorithm used by the authors produces human interpretable association rules of the form:

IF “*severity= non-critical and Time-to-fix = 3 to 30 days and priority= medium*” THEN “*class = doc-bug*” CONFIDENCE: 75.2% SUPPORT: 2.5% (Nayak and Qiu, 2005).

This allows the user not only to use the model to make predictions about projects using the model in the intended manner, but also to examine the association rules created by the model to determine which problems might be the most prominent or impactful and then to adjust the project management strategy accordingly.

The text mining methods set up a large semantic network based on the vocabulary and volume of words in the text. This allowed the authors to extract the most important words and phrases to augment their understanding of the problems encountered in software development, how they are solved and what causes serious delays.

The authors experienced lower-than-expected accuracies in projects in which allocated human resources had been either above or below average. The time required to solve a fault is heavily dependent on the amount of human resources applied to the project. Thus, a project with a large team is likely to solve issues faster than small project teams. However, the authors did not have usable data in their target dataset that reflected the amount of human resources applied to the project. Despite this drawback, the authors were able to extract valuable rules that could aid project managers in estimating the time required to fix problems.

This investigation demonstrates the potential value of data and text mining for project management. The authors show how information about problems encountered, time taken to fix the problems and other common project data can be mined for information to help project management. The authors present a complete data mining process that can be used outside the context of their specific application to extract knowledge from any database.

2.4.2 Case Study 7: Data Mining for the Management of Software Development Processes

Alvarez-Macias *et al* (2004), agreeing with Nayak and Qui (2005), characterised the main problem facing software development projects, and leading to large cost and time overruns, as the difficulty of accurately estimating the cost and duration of the project. Software that was developed to improve the cost and time estimation of software projects have improved

the situation. However, the software requires a number of input variables to be specified. These input variables can be hard to estimate for project managers without years of experience (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).

Alvarez-Macias *et al* (2004) created a data mining application that aimed to replace the current project estimation software by providing project management ‘rules’ that will help both the estimation and management of the project. The data mining process used by Alvarez-Macias *et al* (2004) to achieve their aim is shown in Figure 2-3 and briefly described below.

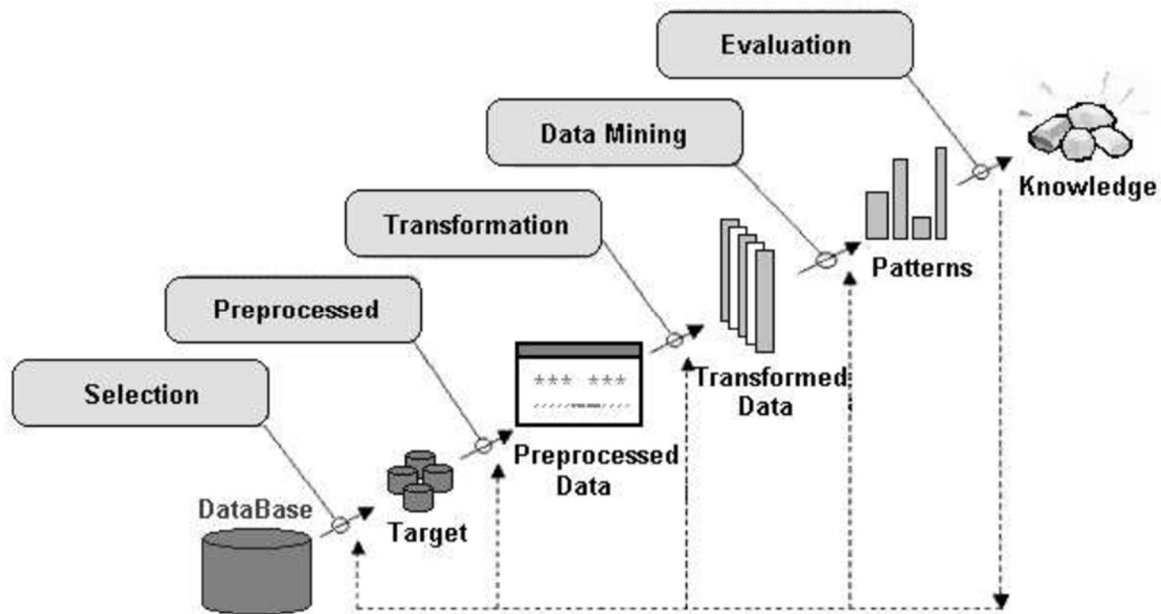


Figure 2-3: Data mining process when applied to software development processes. (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).

- **Selection:** The data that is available is examined for data that could be of interest to the study. The selected data is deposited in a central repository, a ‘data warehouse’.
- **Pre-processed:** The selected data is processed to remove outliers and noise. Absent or missing values are treated.
- **Transformation:** To increase computation efficiency, uninformative and noisy features are removed from the pre-processed dataset.
- **Data Mining:** The transformed and pre-processed dataset is used to induce a set of patterns or rules.
- **Evaluation and Interpretation:** The patterns induced by the model are evaluated by project management experts to determine their validity.

The data collected for the investigation was gathered from project databases and the estimation software. The estimation software input attributes, specified by experienced project managers, were combined with the actual cost and time data from the project databases. Additional information about the number of resources assigned to a project, the specific management policies employed on a project and the general organisation information

was also collected. The collected data was processed to remove noisy features, remove outlier data entries, normalise the data and to impute missing values. These steps were conducted programmatically with close supervision from the expert project managers. The data was split into the input values and the results. The results were used to create class labels for a supervised machine learning algorithm (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).

Two different models were selected to conduct the analysis. The first was an Association Pattern Mining algorithm (See Section 3.5.2). The Association Pattern Mining algorithm is an unsupervised machine learning model that extracts rules from modelling the data, based on the frequency of data features occurring together. The second model was a Rule-Base Classifier (see Section 3.5.5.2). The classifier is a supervised machine learning model that generates human interpretable rules that model the input data to the output class label provided.

The evaluation of the models was done by experienced project managers who inspected the most supported and significant rules generated by the two algorithms. Each algorithm provided several useful rules that software project managers were able to employ to estimate the duration and cost of a project as well as determine the influence of factors such as the number of resources assigned to a problem. These rules enabled less experienced project managers to better quantify their projects and to better balance the cost, duration and quality of a project (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).

The application by Alvarez *et al* (2004) showed how data mining can be used to improve the management of a project. A clear data mining process was presented and applied using two different types of data mining (Association Pattern Mining and Classification). The results of the application were used to balance the required quality of the project with the cost and time demands.

2.4.3 Case Study 8: Data Mining Applied to the Improvement of Project Management:

Balsera *et al.* (2012) presents a complicated and intricate data mining application that is applied to project management in the software development sector. The purpose of examining this paper is not to understand the implementation of the actual data mining application, but rather to examine how and when data mining might be applied to facilitate project management.

Project management is a complex process that involves many interrelated and interdependent internal and external elements that complicate the control of the project. Uncertainty is a large part of project management and it affects many aspects of a project, such as consumption of resources, estimation of time and money as well as the impact of risk on quality (Balsera *et al.*, 2012). The effect of risks and uncertainty have had such a significant impact on project delivery that 31.1% of software project will be cancelled before reaching

completion and 52.7% of software project costing over 189% of their original estimates (The Standish Group, 2014).

Regardless of sector, traditional project management divides a project into 4 main phases, each with an accompanying level of uncertainty (Balsera *et al.*, 2012):

- 1. Initiation:** The project needs are identified and evaluated to determine if the project is feasible and necessary. The uncertainty during this phase is extremely high due to a lack of accurate information. The possibility of errors when estimating is high as each project is unique with its own complexities and requirements that might not yet be known.
- 2. Planning:** The planning phase breaks down the problem into more detailed activities and attempts to develop a solution. This reduces estimate uncertainties as more detailed information becomes available.
- 3. Execution:** Once the tasks of the project have been clearly scoped and defined, the execution of the project can begin. Monitoring and adjusting techniques are used to maintain control of the project. During the execution phase the inherent uncertainty in the project reduces dramatically but the possibility of incorrect estimates impacting the project still exists. These incorrect estimates can have a severe impact because large changes are not possible at this point.
- 4. Closure:** The closure phase of the project is when uncertainties are reduced to zero. The project is handed over and data is collected on the performance of the team, errors made, delays encountered, and the extent to which the project was a success. This forms part of the 'lessons learned' document that will be used to help future project managers realise opportunities and avoid mistakes.

Data mining is typically used to help reduce uncertainties and provide a reliable source of information on which decisions can be based. As such, the project phases in which data mining will have the most impact are the pre-planning and planning phases when accuracy information is often scarce or difficult to acquire. (Balsera *et al.*, 2012).

In order to provide usable estimating information during the planning phase, the data mining application will need to contain information from every phase of the project (Balsera *et al.*, 2012). Some issues can arise during the data collection, as every project is different and the data types, fields or indicators stored may vary. Effort is needed to generate a usable target dataset for data mining from varying sources and of varying levels of data granularity (Balsera *et al.*, 2012).

This investigation shows the value of data mining project information assisting project management. The ability of data mining to reduce uncertainty and to provide accurate, data-driven knowledge for use in decision making is most valuable during the initiation and planning phases of the project. Every project, regardless of sector, follows the same four basic phases. This shows that data mining project information in the construction sector could

aid project management if it is applied during the initial phases of a project. Data mining should be based on previous completed projects to help provide realistic data to reduce uncertainties during the early project phases.

2.5 Construction Project Environment:

In order to apply the same data mining processes used to assist project management in the case studies to the construction sector, it must first be established that construction projects have the same phases, similar uncertainties and that reducing the uncertainties within the phases will improve the management of the project. This section aims to show that construction projects follow similar phases and that the uncertainty within construction projects are comparable to those in a software development project. The uncertainties within a construction project are based around the critical success criteria. Previously, these criteria were known as the 'Iron Triangle' of project management: i.e. Cost, Time & Quality. In this section, the more contemporary sustainable criteria are discussed and adopted, i.e. Cost, Time, Quality, Environmental Impact and Social Impact.

2.5.1 Construction Project Management Phases:

According to Balsera et. al. (2012), all projects, regardless of sector, progress through the same four basic phases: initiation, planning, execution, and closure. However, this four-stage view does not consider the entire project lifecycle. The comprehensive project lifecycle is typically divided in to 5 phases. The 5 phases shown below are adapted from Archibald, Di Filippo and Di Filippo (2012) and the PMBOK as presented by Nicholas and Steyn (2012):

1. **Project Initiation Phase:** This first phase of a project is where the project idea is generated, or the specific needs are identified. The project idea is documented, and the feasibility of the project is assessed. The objectives of a project are weighed against the cost and timeline of the project and compared with the available resources to determine if the project is sound. Projects that are deemed to be feasible will be assigned to a project team within the organisation that conceptualised the project.
2. **Project Planning Phase:** The planning phase of the project is an important phase where the project scope and deliverables are set. The work packages for achieving these goals should be detailed and well planned. This plan should guide the project team through the procurement of all the necessary services, materials, of funding and any other required resources. The project plan should detail how the project will be delivered and how the team will handle less tangible aspects of the project: such as how to handle project risks; how to communicate with stakeholders and how to manage supplies and the project.
3. **Project Execution and Monitoring Phase:** The execution phase of the project is where the project deliverables are generated. The methods of achieving the deliverables and how to handle any situation that arises during the execution of the project should be derived from the project plan. The execution phase of the project is most commonly associated with project management as it is where the project

management team can allocate resources, control work flows, and manage communication between project parties. The project team uses key performance indicators to monitor the execution of the project and the work packages to prevent scope creep. They track deviations in cost and time from those specified in the project plan.

4. **Project Handover and Operation:** Once the project has been executed, the project is handed over to the client for operation. This is often where many organisation's involvement with the project will end. In such a case they will not have an operation phase but instead move directly to project closure and evaluation. The operation phase for civil infrastructure is typically several decades long. To ensure the continued use of the asset, the operating organisation is required to perform all necessary maintenance and improvements throughout the usable lifetime of the asset.
5. **Project Closure and Evaluation:** This phase represents the end of the project. This specific stage could occur for certain project partners at the handover of the project. For example, an organisation that supplies materials to a construction site will not have a continued part in the operation phase of the project. Therefore, their closure and evaluation phase will happen at the project handover. The closure phase is when all parties' contracts are terminated. Typically, a post-mortem investigation/discussion will be held to determine the success and failures of the project, in order for the project members to learn from their mistakes. All final project documentation will be delivered and stored.

Each of these project phases are accompanied by their own level of uncertainty and ability to influence the project (Archibald, Di Filippo and Di Filippo, 2012; Nicholas and Steyn, 2012). As shown in Figure 2-4, the ability of stakeholders to influence the project, the risk associated with the project, and the uncertainty in construction projects decrease as the project progress, however the cost of changes rises with time.

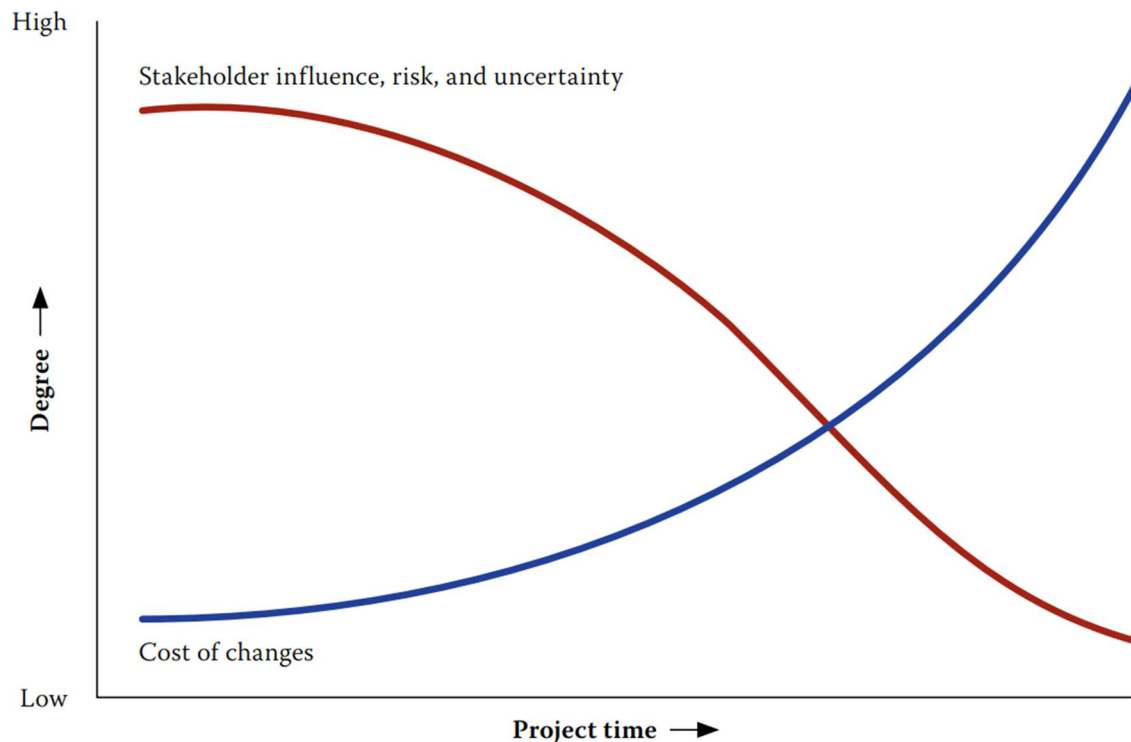


Figure 2-4: Impact of variables and uncertainty based on project time (Schoonwinkel, Fourier and Conradie, 2016)

The ability to reduce the risk exposure of the project, or to reduce the uncertainty of the project during the initial stages of the project will prove to be the most valuable, as the cost of implementing any changes will be significantly lower than later in the project. The uncertainty within a project can be reduced by providing more accurate estimates, more accurate predictions of resources required and by providing clarity in any meaningful way (Balsera *et al.*, 2012; Schoonwinkel, Fourier and Conradie, 2016).

2.5.2 Sustainable Project Management Success Criteria:

Project management has, since its official promulgation in the 1950s, attempted to set out criteria against which project success can be measured. Traditionally these have been the so-called 'Iron Triangle' of Cost, Time, and Quality. Project managers have traditionally focused on balancing the competing factors within a project and to deliver projects on time, on budget and of required quality. A potential drawback of placing these three values at the centre of project success is that other important success factors could be ignored or missed. These success metrics seem to ignore customer satisfaction and other external impacts of the project (Caccamese and Bragantini, 2012; Ebbesen and Hope, 2013).

The impact of a business, project or development on the environment and society can no longer be ignored. Therefore, the adoption of sustainability has seen a surge in recent years. In the business world the traditional 'bottom-line' of monetary value is being challenged by the concept of the 'triple bottom-line' of Economy, Society, and Environment (Tharp, 2012; Ebbesen and Hope, 2013).

In the project sphere, several alternative project success criteria have been proposed to incorporate the goals of sustainable development. The most well-known and accepted version is the so-called ‘sustainable project management star’ or ‘sustainable project management circle’ shown in Figure 2-5.

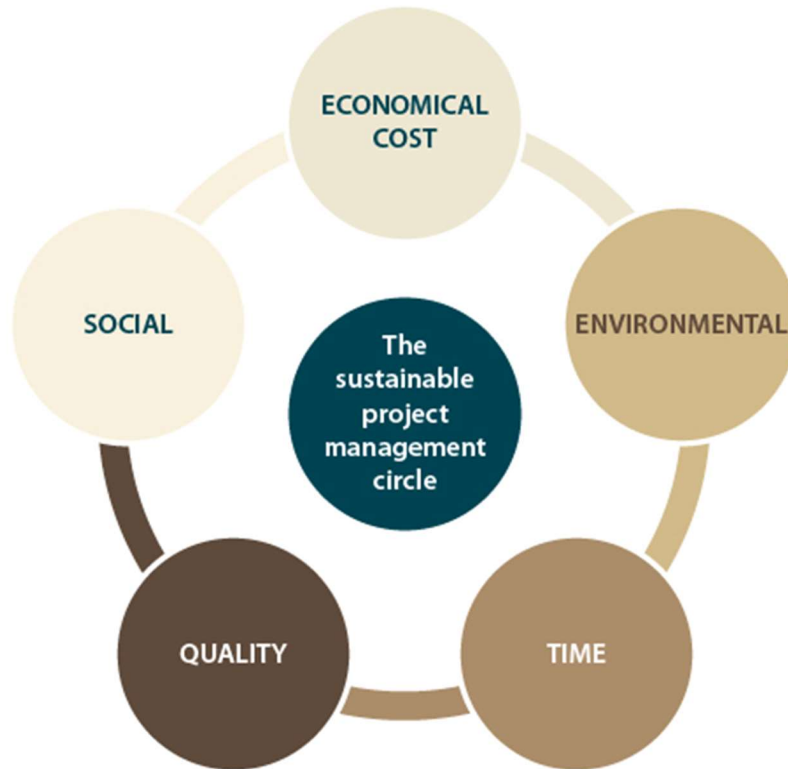


Figure 2-5: Sustainable Project Management Circle:

These 5 criteria are linked in the same way as the three in the ‘Iron triangle’. If the project aims to reduce its environmental impact and increase its positive social impact, the trade-off could be in increased cost or time taken to complete the project. The project manager must therefore balance the cost of the project, the time taken to complete the project, the required quality, the environmental impact of the project and the social impact of the project (Tharp, 2012). These five success criteria will be adopted as the success measures for project management in this research.

2.6 Discussion of the Literature:

The case studies of data mining applied to the construction sector are summarised below. The suitability of applying data mining to construction projects is discussed using information gathered from the investigation of the construction project environment and the software development project environment. Finally, using the acquired knowledge, a data mining process for applying data mining to the construction sector to facilitate project management is synthesised.

2.6.1 Summary of Construction Data Mining Case Studies:

Five case studies of data mining applied to the construction and civil engineering sector were examined. These examples provided valuable information that is summarised and discussed below. The impact of the information gained from the examples is discussed later, after the data mining processes presented by the case studies are examined.

Presented in Table 2-1 are the project goals, the data used in each project and additional notes about the process or other important factors.

Table 2-1: Data Mining in the Construction Sector Case Study Summaries:

Project Goal:	Data Used:	Additional Notes:
Residual Value Assessment of Heavy Construction Equipment (Fan, Abourizk and Kim, 2008).	Numerical and Categorical data obtained from web databases.	<ul style="list-style-type: none"> The investigation had a well-defined goal. The goal was re-examined, and the scope changed to account for data insufficiencies and environment complexities. The validation of the data was emphasised along with the importance of the usability of the application.
Using Data Mining to Discover Knowledge in Enterprise Performance Records (Lee, Hsueh and Tseng, 2008).	Text, categorical and numerical data from maintenance databases, enterprise records, design documentation and construction documentation.	<ul style="list-style-type: none"> The investigation had a well-defined and specific goal. The authors followed a definite data mining process. The process applied was both linear and iterative in nature. The authors showed that the output from the model might require human interpretation before being usable.
Data Mining Tender Bid Information to Evaluate the Best Bid Selection Policy (Chaovalitwongse <i>et al.</i> , 2012).	Numerical and categorical data from web databases.	<ul style="list-style-type: none"> The investigation had a well-defined goal, to evaluate bid selection policies. The authors conducted a preliminary investigation of the data to ensure that the data is valid and not extremely skewed. The applicated used ratios and other highly transformed data in the target dataset.

Project Goal:	Data Used:	Additional Notes:
Predicting Construction Cost Overruns Using Data Mining (Williams and Gong, 2014).	Numerical and text data from government agency databases.	<ul style="list-style-type: none"> The investigation had a definite goal, to predict if a cost overrun will occur. The goal of the application was very ambitious. Insufficient data was collected to model the problem, which resulted in low prediction accuracies. The authors demonstrated the use of combining numerical and text data to provide a more complete and representative target dataset.
Data mining to detect early warning signs of project failure by mining unstructured text from site meetings (Alsubaey, Asadi and Makatsoris, 2015).	Text data from project documentation (site meeting minutes)	<ul style="list-style-type: none"> The investigation had a definite goal, to detect construction risks from site meeting minutes. Obtaining good results by only using text data shows that data contained in project documentation can be extremely useful.

The large range of applications, data sources used, data mining models used, validation techniques used, and the different processes applied shows that data mining in the construction sector is not restricted to only a small part of the sector.

2.6.2 Suitability of Applying Data Mining to Project Management in the Construction Sector:

The suitability of applying data mining to the construction sector with the specific aim of aiding project management is an important aspect that must be considered. To determine if construction projects are suitable for data mining, four questions need to be asked and answered:

1. Do construction projects progress through the same phases as software development projects?
2. Does the uncertainty within construction projects follow the same curve as the uncertainty with software development projects?
3. Would reduction of the uncertainty within the phases of a construction project facilitate project management?
4. Do construction projects have stores of data large enough to warrant the use of data mining?

Project Phase Similarity: Project management in the software development sector follows 4 basic phases:

1. Initiation.
2. Planning.
3. Execution.
4. Closure and Evaluation.

As show in Section 2.5, construction projects follow similar phases:

1. Initiation.
2. Planning.
3. Execution and Monitoring.
4. Handover and Operation.
5. Closure and Evaluation.

These five phases in the life-cycle of construction projects are very similar to the phases in software development projects. The main difference between the two is that construction projects have a dedicated operation phase. The operation and support of the software, after being delivered to the client, is included in the execution phase of software development projects. This difference can be attributed to the fact that software projects tend to be developed by one organisation, either as a product or in contract to a client. The software development team will work on the project during its development and will retain ownership of the project during its maintenance and upkeep (software updates and bug fixes).

In contrast, construction projects typically separate the execution and operation phase due to the large number of parties that end their interaction with the project once the asset has been delivered. A different set of parties are typically contracted during the operational phase. For example, material suppliers, contractors and engineers that were part of the execution phase do not typically partake in the maintenance and operational phases.

The length of the operational and maintenance phase of construction projects also contributes to the separation of the execution and the operation and maintenance phases. The clear boundary between the construction of the asset and the operation of the asset is exacerbated by the length of the operation and maintenance phase and contributes to the reason for construction projects having dedicated operational and maintenance phases.

This difference does not seem to be significant enough to necessitate an entirely different formulation of the data mining process and how it can be utilised to facilitate project management. Furthermore the data mining will typically be applied during the initiation and planning phases, which are identical for software projects and construction projects.

Project Uncertainty Levels Similarity: In Case Study 8, Balsera *et al* (2012) states that the uncertainty and risk exposure during the initiation phase is high with it gradually reducing to zero as the project progresses through its 4 phases. In Figure 2-4, Schoonwinkel, Fourier and Conradie (2016) present a similar scenario for construction projects. The acceptability of

utilising the same data mining methodology for construction projects as was used on software project is increased by the similar levels of uncertainty and risk exposure throughout the project phases.

Reduction of Uncertainty to Aid Project Management: The sustainable project success criteria (Cost, Time, Quality, Environmental Impact, and Social Impact) each have their own uncertainty at each stage of a construction project. The overall uncertainty curve within a project comprises the combined uncertainties for each of these success criteria.

Balser *et al* (2012) stated that the reduction of uncertainty, especially during phases of high uncertainty, is the overall goal of data mining to facilitate project management. The type of uncertainty (Cost, Time or Quality) does not influence the overall methodology or data mining process. Since the specific success criteria has no influence on these factors, it is reasonable to conclude that the addition of the sustainable project criteria will not require the data mining methodology used for software projects to be altered for use in the construction sector. Therefore, reducing the uncertainty of any of the five sustainable development success criteria during any construction project will facilitate the management of that project.

Sufficient Data: The construction sector produces a large amount of data and documentation throughout the 5 phases of a construction project. The data produced is stored in a variety of ways and formats, such as in databases, spreadsheets, documentation etc. As shown by the variety of data mining applications and the many different sources of data in Section 2.3, the sector is sufficiently data-rich to make data mining a viable project management aid. The main challenge to the sector is acquiring the required data for a data mining application.

Capturing data from project documentation that is stored in a physical format (books, papers etc.) poses a significant problem. Other problems, similar to the problems encountered by Lee, Hsueh and Tseng (2008), such as scattered, or poorly recorded data could well be a common theme in the construction sector in South Africa. However, larger engineering firms, client bodies, and construction firms are likely to have data stored in project databases to keep track of their projects. These project databases should prove to be a valuable data source for data mining.

Conclusion: The suitability of applying data mining to construction projects to facilitate project management is an important consideration. The combination of the past successes found by applying data mining to project management in the software development sector, the similarities between software projects and construction projects (both in terms of phases and uncertainty), and the amount of data stored within the construction sector strongly suggests that data mining is highly suited to assisting project management within the construction sector.

The importance of good quality and accurate data to use as the basis for data-driven decision making cannot be overstated. Transitioning the South African construction sector to a data-driven decision process based on data mining could increase the profitability of its construction companies. Having more and better information will allow all parties in the

construction sector to reduce their costs, decrease their delivery times, meet quality demands, reduce their environmental impact, and create long-lasting and sustainable growth for the sector.

2.6.3 Definition of a Data Mining Process for Application in Construction Projects:

The definition of a data mining process that can be used by any investigator into construction projects is part of the goal of this investigation. The case studies of data mining applied to both the construction sector and the software development sector each have their own application specific data mining processes.

Knowledge about the data mining process from the case studies is extracted and used to synthesise a data mining process. The data mining process should act as a guiding document for novice data mining practitioners to create a data mining application that will help them produce information that can be used in the management of construction projects.

2.6.3.1 Knowledge from Case Studies with Defined Data Mining Processes:

Three case studies that explicitly defined their data mining process, are presented in Table 2-2:

Table 2-2: Comparison of Data Mining Processes

	Case Study 2:	Case Study 6:	Case Study 7:
Case Study:	Using Data Mining to Discover Knowledge in Enterprise Performance Records (Lee, Hsueh and Tseng, 2008).	Data Mining Application in a Software Project Management Process (Nayak and Qiu, 2005)	Data mining for the management of software development processes (Alvarez-Macias, Mata-Vazquez and Riquelme-Santos, 2004).
Data Mining Process:	<ol style="list-style-type: none"> 1.) Data Selection 2.) Data Cleaning and Preparation 3.) Data Reduction and Coding 4.) Algorithm Selection 5.) Mining and Reporting 	<ol style="list-style-type: none"> 1.) Data Acquisition 2.) Data Pre-processing <ol style="list-style-type: none"> a. Defining Goals b. Field Selection c. Data Cleaning d. Data Transformation 3.) Data Modelling and Mining 4.) Assimilation and Analysis of outputs 	<ol style="list-style-type: none"> 1.) Data Selection 2.) Data Pre-processing 3.) Data Transformation 4.) Data Mining 5.) Evaluation

Although the data mining processes used by the three case studies do vary, there are many similarities. The data mining process presented in Table 2-2 are discussed below, and the knowledge gained will be combined with knowledge from the other case studies to synthesise a full data mining process.

Case Study 2: The process used to discover knowledge in enterprise performance records is a relatively complete process. Due to insufficient information in the maintenance databases, a significant amount of energy and time was expended in compiling a target dataset from many sources. The authors manually supplemented the maintenance database data with design and construction information. This is reflected in their data mining process by the three steps: 1) data selection 2) data cleaning and preparation and 3) data reduction and coding. These are essentially different sub-steps of Data Pre-processing. The ‘algorithm selection’ step can also be viewed as a sub-step of the ‘mining’ step.

Case Study 6: The data mining process defined in the case study, ‘Data Mining Application in a Software Project Management Process’ is a general process that is not application-specific. Defining the acquisition of the data as a separate step at the start of the data mining process is a clear indication by the authors that careful attention must be paid at this stage to enable the application to accurately model the system. This expanded data pre-processing step illustrates the authors’ awareness that the pre-processing step may be extensive and require significant amounts of work. The inclusion of the goal definition sub-step in the pre-processing step is also an important insight.

After data was acquired, the investigators set a specific goal that was achievable with the data they had acquired. Having a well-defined goal before commencing pre-processing ensured that the data assembled into the target dataset would contain the necessary information to model the system accurately and present data that would fulfil their goal. The importance the authors assigned to the evaluation of the results obtained from the data mining model is shown by the fact that they dedicated an entire step specifically to assimilating the results and evaluating them.

Case Study 7: The data mining process used in the third application lacks some information and key details presented in the other applications. The investigation had a definite goal, but the data mining process is missing a goal definition step or sub-step.

Data selection and data transformation is often considered to be part of the Pre-Processing step, as illustrated in Case Study 6. Case Study 7 also lacks a definite data acquisition step. These shortcomings are understandable since the authors did not set out to define a data mining process as such. However, the separate evaluation step reinforces the message that evaluating the results from a data mining process is critical to the success of the project.

2.6.3.2 Knowledge from Case Studies without a Defined Data Mining Process:

The case studies without a defined data mining process, such as those summarised in Section 2.6.1 may nevertheless contribute significantly to the final data mining process.

Goal Definition: As can be seen in Table 2-1 in Section 2.6.1, the case studies of data mining applied to the construction sector all had clearly defined goals at the outset of the data mining process. This well-defined goal informed the collection of the data, the pre-processing methods used, the data mining algorithm selected, the validation steps employed and the format of the results of the data mining. The importance of setting a clear goal cannot be underestimated and warrants the introduction of a goal definition step at the start of the data mining process.

Data Acquisition: Having accurate and complete data from which to formulate a target dataset is vital to the success of a data mining application. Ensuring that the target dataset contains the necessary information to produce an accurate model of the system will greatly increase the accuracy of a data mining exercise. The importance of acquiring data can be seen by the fact that all the authors of the case studies discussed above discuss the origin of the data used in their application and how they retrieved it. This warrants the inclusion of a step in the data mining process specifically devoted to acquiring data.

The data used for construction case studies was collected from a large range of sources. Amongst them were project databases, online databases, company spreadsheets, construction and design documentation, and project documentation. All of these sources should be used when acquiring data for data mining in the construction sector.

Data Pre-processing: The pre-processing of the collected database into a target dataset that will be used for the data mining application is mentioned in every case study examined. The pre-processing of the data varied from selecting and extracting useful features to cleaning, reducing and transforming the data.

It is clear from all the case study examples that a pre-processing step must be included in the data mining process. For the purpose of this study, the pre-processing step will be considered to include all the data processing of the data to produce the required target dataset. This is done so as to ensure that each step in the data mining process is a clear and separate process.

Mining and Modelling: The actual application of the data mining algorithm in the case studies tends to be described through the authors' explanation of choice of models, how they were applied, and the results obtained from the models. The selection of an appropriate model for the specific application is an important part of ensuring that the data mining application achieves the desired results. However, the authors do not typically discuss the large variety of types of data mining that exist (Clustering, Classification etc.) and the many algorithms that fall under these different types (Decision Tree classifier, Naïve Bayes classifier etc.).

The selection of the appropriate algorithm and applying it correctly can be a complex step that requires knowledge of the available data mining types, models and what each can be

used for. Therefore, a step in the data mining process is reserved for the process of selecting models that are appropriate for the application and applying them to the target dataset.

Validation and Evaluation: In each of the case studies discussed the results obtained from the mining and modelling step are almost always verified in some way. The actual validation and evaluation processes used to determine if the model produced accurate, reliable results will differ between applications and data mining types, depending on what is appropriate. The importance of evaluating the results is emphasised by the majority of case study authors. Therefore, a step in the data mining process will be devoted to the validation and evaluation of the results obtained from the data mining algorithm.

Linear and Iterative Nature: Several case study authors described the data mining process as both a linear and an iterative process. The data mining process definitely flows in a linear manner from the collection and pre-processing of the data to the modelling and validation of the data. However, at the end of each of these steps some of the authors returned to a previous step to alter the pre-processing applied to the data or the selected model.

This iterative process should be driven by the desire to achieve the goals set for the application with results that are as accurate as possible while ensuring that the model remains usable and cost effective.

2.6.3.3 Data Mining Process:

Knowledge gained from the case studies that explicitly define a data mining process as well as that from the case studies without such a process was combined to produce a process that may be used to apply data mining to improve project management in the construction sector (Figure 2-6). The data mining process synthesised here is discussed in Chapter 3. Each step within the process is individually addressed and its goals and methods discussed.

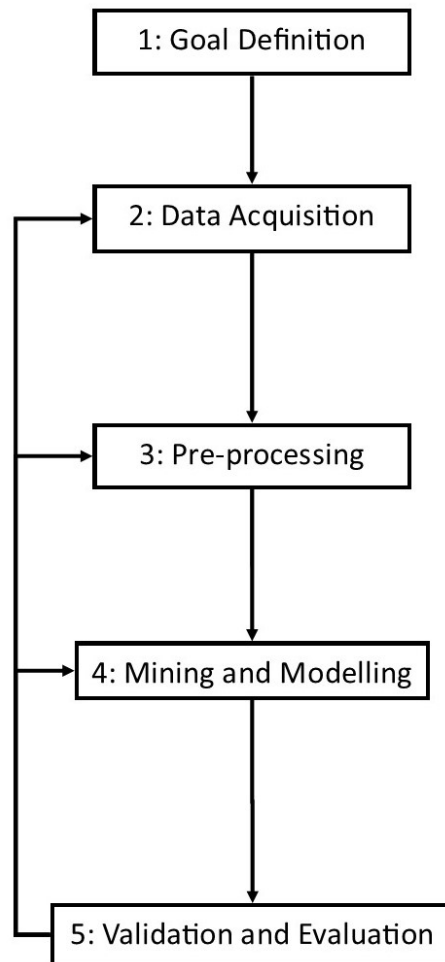


Figure 2-6: Synthesised Data Mining Process

2.7 Conclusion:

The aim of the literature review in Chapter 2 was to synthesise a data mining process that can be used to mine the information available in the construction sector in order to improve the management of construction projects. The 5 case studies of data mining in the construction sector shows that the construction sector has sufficient stores of data to allow for data mining techniques to be applied. The 3 case studies of data mining in the software development sector and the information about construction project phases revealed how data mining can be used to improve the management of construction projects. The large amount of available data, the known methods for improving data mining and the data mining processes used in the 8 case studies allowed for a data mining process to be formulated specifically aimed at improving the management of construction projects. A summary of Chapter 2 is provided in Appendix 1.

3 Data Mining Process:

3.1 Introduction and Background Information:

To fulfil Objective 2 of Aim (see Section 1.2), Chapter 3 discusses the data mining process synthesised in Chapter 2. The data mining process outlined in Chapter 2 was combined with information from four leading data mining textbooks:

- **Data Mining and Analysis: Fundamental Concepts and Algorithms** by Zaki and Meira (2014)
- **Data Mining; the Textbook** by Aggarwal (2015).
- **Data Mining: Concepts and Techniques** by Han, Kamber and Pei (2012).
- **Data Mining: Practical Machine Learning Tools and Techniques** by Witten and Frank (2005).

These authors are highly respected in the field of data mining and machine learning and their textbooks are used world-wide in both undergraduate and postgraduate data mining contexts.

The authors provide highly detailed discussions about the different types of data mining, the algorithms in each type, and the implementation of each algorithm. This level of detail is beyond the scope of this investigation. This chapter will simply provide an overview of the data mining process, the types of data mining and some of the methods. The sections of this chapter follow the data mining process. As such, the data mining process and an overview of the sections within this chapter are presented in Figure 3-1.

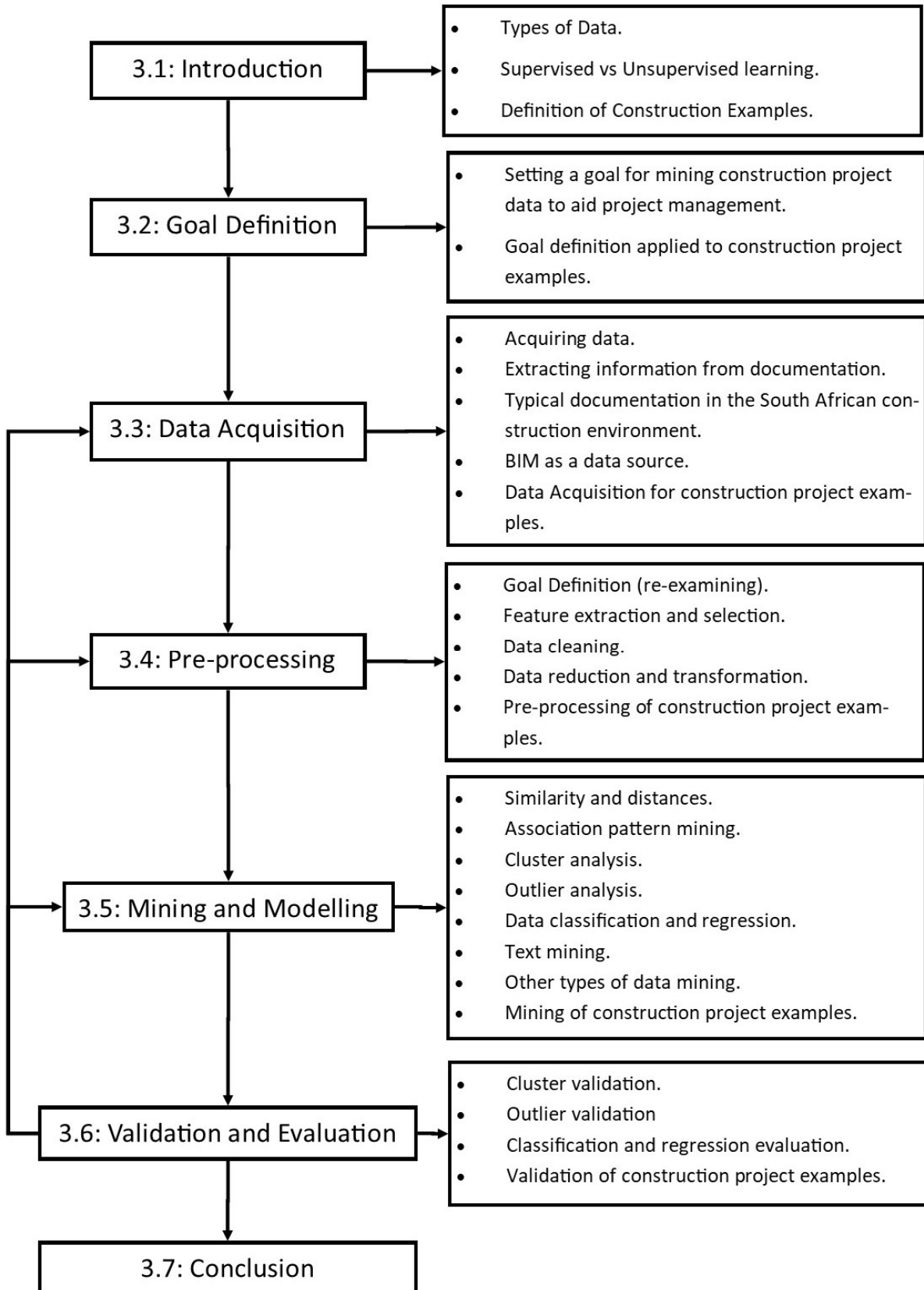


Figure 3-1: Data Mining Process and Chapter 3 Overview

Each step in the data mining process will be addressed and discussed in sufficient detail to enable a novice data mining practitioner to understand the process. There is a brief discussion

of the types of data and the difference between the two main types of machine learning (supervised and unsupervised) in this section.

In order to tie the data mining process to the construction industry, two example problems faced by the South African and Global construction industries are defined in Section 3.1.3. After each step within the data mining process has been discussed, the impact of that step on the construction project examples is briefly discussed. This chapter discusses the data mining process without providing any data mining implementation resources, which are presented and discussed in Chapter 4.

3.1.1 Types of Data:

The data mining methods and algorithms chosen to examine a dataset are dependent on the format of the data that will be mined. Therefore, it is important that the investigator is familiar with the type, content and format of the data. This section looks at the difference between structured and unstructured data, the two types of data typically collected during the data acquisition phase. The difference between dependency and non-dependency orientated data is explained, along with some considerations that must be made before selecting a data mining model.

Structured vs Unstructured Data: Structured, unstructured or semi structured data refers not to some characteristic of the data but rather to the level of organisation and format of the data. Structured data refers to data with a high degree of organisation, typically data stored in a data model. This data can be easily interpreted and used by both a user and a computer system (Grimes, 2007).

Unstructured data is data that does not have a predefined data model nor is it organised in a pre-defined manner. Unstructured data typically refers to text, which can easily be understood and used by humans but is typically not usable by a computer systems (Grimes, 2007).

Semi-structured data falls between structured and unstructured data and typically refers to tables and lists of data contained within text documents. This data is therefore easily used by humans but cannot be used directly by a computer system (Grimes, 2007).

Nondependency-Oriented Data: Nondependency-oriented data can be considered as the simplest form of data used in data mining. The dataset consists of records also referred to as data points or feature-vectors, depending on the application. Each of these records contains a set of data fields, also referred to as attributes or features (again, depending on the application). The nondependency-oriented aspect of the data refers to the fact that no direct or indirect internal relationship exists between data fields within the same record (Aggarwal, 2015). For example: the age, gender, postal code and marital status of a person have no direct internal relationship.

Nondependency-oriented data can be sub-divided into quantitative multidimensional data, categorical or mixed attribute data, and binary data (Han, Kamber and Pei, 2012; Aggarwal, 2015):

- Quantitative multidimensional data are attributes that are continuous, numeric and have a natural ordering, for example a person's age. Quantitative multidimensional data is the most common form of data and for which most data mining techniques are developed.
- Categorical or mixed attribute data refers to discrete unordered attributes. Attributes such as gender and postal code have discrete values with no natural ordering. Mixed-attribute data refers to data that contains a combination of categorical and numeric data.
- Binary data can be considered as a special case of either multidimensional quantitative data or multidimensional categorical data. Multidimensional quantitative data because an ordering exists between the two values and multidimensional categorical data because each value can assume one of two discrete values.

Dependency-Oriented Data: As discussed above, nondependency-oriented data can be assumed to have no direct or indirect internal relationships. Dependency-oriented data records contains attributes that have internal relationship between the entries. Since data mining is focused on finding relationships between data attributes, it is important that all data relationships (dependencies) are known beforehand. These relationships may be temporal or spatial in nature or defined through some other pre-existing relationship.

The internal dependency between data attributes can either be explicit or implicit. Explicit dependencies are dependencies that are definitive and unquestionable, such as network data or graph data where the edges define explicit dependencies. Implicit dependencies are dependencies that are not defined in such a rigorous manner but are known to generally exist with the data from certain domains.

Dependency-oriented data can be sub-divided into time-series data, discrete sequences, spatial data, and network or graph data (Han, Kamber and Pei, 2012; Aggarwal, 2015):

- "Time-series data contain values that are typically generated by continuous measurement over time" (Aggarwal, 2015). For example, a temperature probe or rainfall measurement probe both measure a value over a period of time. Since the data is more likely to vary smoothly over time, the attributes have an implicit relationship. This implicit relationship must be noted and used in the data mining application, as large sudden jumps in the measurement could be points of interest or a cause of concern. It is possible to generate a multidimensional time series dataset, by combining several time series datasets over the same period with the same time step.
- Discrete sequences are nearly identical to time series data. They only differ in the attribute collected. The time series measures a continuous measurement whereas the discrete sequence measures a categorical value, such as the login status of a computer over time. The computer can either be logged in or logged out at any point in time, but this can change as different people use the computer.

- Spatial data is data that has been collected at different spatial locations. The attribute recorded at each spatial point is explicitly related to the data collected at every other spatial point via the spatial distribution or the measuring network. For example, meteorologists collect the surface temperature of the oceans to predict weather behaviour. The temperature reading at each point is related to the temperature reading at the spatial location adjacent to it. The attributes at each spatial location can be collected over a period of time, allowing the creation of spatiotemporal data. Spatiotemporal data have two explicit dependencies, spatial and time.
- In network and graph data the recorded data values correspond to a node within the network. The edge of the network usually defines the relationship between the data values. The fact that a network representation of data is a generalised form, increases its usefulness when applying data mining. Multidimensional data can be converted to network data by assigning each record to a node. The data can then be mined to determine the edges/relationships. Network data is extensively used in data mining social networks, web graphs and chemical compound databases.

The form of data most typically used within the construction sector is nondependency-oriented data such as the cost, duration and scope of a project.

3.1.2 Supervised vs Unsupervised Machine Learning:

Two key concepts in data science and machine learning are supervised learning and unsupervised learning. Supervised machine learning is where the output variable (Y) and in the input variable (x) are known and an algorithm is used to learn the mapping function (f) of the input variable (x) to the output variable (Y).

$$Y = f(x)$$

The algorithm iteratively learns how to map the input variable to the known output variable until it has reached an acceptable level of success, almost like a teacher with the answer teaching a student. Therefore, it is known as supervised learning. Supervised learning can produce discrete or continuous outputs which are used in classification and regression respectively (see Section 3.5.5) (Han, Kamber and Pei, 2012).

Unsupervised machine learning is the process in which the input variable (x) has been provided but there is no correct corresponding output variable (Y). Unsupervised machine learning algorithm aims to learn and model the underlying structure and distribution of the dataset in order to assign each data point a value, based on their similarity. Since there is no correct answer and no teacher it is known as unsupervised machine learning. These algorithms can be used for clustering, outlier detection and association rule making (see Sections 3.5.3, 3.5.3 & 3.5.2 respectively) (Zaki and Meira, 2014).

3.1.3 Examples of Data Mining to Facilitate Construction Project Management:

This section introduces two examples of data mining in the construction sector for the facilitation of project management. These examples will be used to aid the explanation of the data mining process by demonstrating how each step may be utilised. One example is a real example based on an investigation by Lee *et al* (2011) into predicting construction project cost overrun. This investigation was selected for its clear demonstration of the data mining process and for its useful results. The second example is a fictitious example based on the risk of on-site accidents and injuries in the construction industry, identified by Akintoye and MacLeod (1997), Fang *et al* (2004) and Tang *et al* (2007).

Data Mining to Predict Cost Overrun: As stated by Williams and Gong (2014), the global construction industry suffers from project cost overruns. These cost overruns can be caused by factors such as: delays in payment by the client; financial difficulties experienced by the contractor; mistakes and discrepancies in design documentation; lengthy bureaucracy and many other factors (Niazi and Painting, 2017). The effects of cost overruns are wide-reaching and may extend beyond the project and the contracted parties.

Using data mining techniques, Lee *et al* (2011) was able to produce a predictive model that was able to forecast the likely cost variance of a construction project in South Korea. While facilitating project management was not the specific goal of the application, it is clear that the information predicted by this model could be used for this purpose by confirming the initial cost estimates and indicating which projects are likely to encounter problems. The project managers would then be able to dedicate their resources to the required areas to help prevent the cost overruns from materialising (Lee *et al.*, 2011).

Due to the clarity of the process used by Lee *et al* (2011), their application is useful as a practical example for the explanation of the data mining process described in this chapter. After the aim and methods of each step in the data mining process has been described, this example is revisited to discuss how Lee *et al* (2011) applied the specific step to their data mining application.

Data Mining to Predict the Impact of On-site Accidents: On-site accidents and worker safety are consistently ranked as some of the greatest project risks faced by the global construction industry. The impact on a project of an on-site accident will depend on the specific legislation of the country of operation, but it typically negatively affects the timeline of the project and could financially impact the construction company through fines and legal sanctions (Akintoye and MacLeod, 1997; Fang *et al.*, 2004; Tang *et al.*, 2007).

Risk management is the process of identifying, analysing, monitoring and initiating responses to project risks by use of a variety of specially designed techniques. The aim of risk management is to increase the efficiency and value of project delivery (Tang *et al.*, 2007). The identification and analysis of construction project risks requires accurate information on the likelihood of a risk materialising, as well as the impact that the risks will have if they do

materialise. This fictional example presents a possible data mining application that is designed to estimate the impact of an accident on the duration of a project.

3.2 Goal Definition:

Goal definition, being the first step in the data mining process is an important part of a data mining project. Having a clearly defined and reachable goal will inform the investigation as to the specific data to collect, the format of the output from the model and how the results will be utilised. This in turn influences the selection of data mining models and the required level of accuracy. This initial goal definition stage is included in the data mining process to prime the investigation with a broad scope.

3.2.1 Setting a Goal to Facilitate Construction Project Management Through Data Mining

Construction projects are often complex, multimillion-rand endeavours that have multiple parties that must be coordinated to ensure that a shared goal is reached. In order to formulate a clear and concise goal for a data mining application in the complex and unique construction environment the investigator needs to be familiar with the environment. A competent investigator will be aware of the organisation's role in construction projects and its goals in executing the project, be it a construction company, an engineering firm, a government body or a private client entity. Cognisance of these aspects will enable the investigator to establish a clearly defined goal that will yield useful results.

As discussed in Section 2.5, construction projects follow the same basic project phases as any other project: initiation, planning, execution and monitoring, handover and operation, and closure and evaluation. Depending on the role of the organisation in the project and the organisation's goals for the project, the investigator attempts to reduce the uncertainty within a specific stage by providing useful information gained from data mining previous construction projects. In order for the data mining application to be successful, uncertainty should be reduced in at least one of the five sustainable project success criteria i.e. Cost, Time, Quality, Environmental Impact, or Social Impact.

The investigator should aim to provide useful information about at least one of the five success criteria of sustainable project management during at least one of the five stages of a construction project by using one of the specific types of data mining types discussed in Section 3.5. Rather than aiming to provide total clarity about the entire project, the investigator should attempt to produce a system that can repeatedly provide useful information.

Balsera *et al* (2012) states that data mining should be applied during the initiation and planning phase as the inherent uncertainty within these phases are high. Pospieszny (2017) argues that useful information can be obtained by applying data mining during any of the 5 phases of a project. Both of these approaches are valid as the key factor is the reduction of uncertainty. The reduction of uncertainty could be more impactful during the initiation and

planning phase, but every phase has some degree of uncertainty and by reducing it the management of the project can be supported.

A second goal definition step is included in the pre-processing stage. This step of the data pre-processing phase is included so that the investigator can re-examine the initial goal in the light of the available data.

3.2.2 Goal Definition Applied to Construction Project Management Examples:

The discussion of the data mining examples defined in Section 3.1.3 is continued in the section where the Goal Definition step is applied to both the examples.

Goal Definition for Prediction of Cost Overrun Example: The goal of the data mining application developed by Lee *et al* (2011) was to provide estimates of the financial success, or failure, of a project. The goal of the project was not specifically to aid project management, but it is clear how knowing if a project will have a cost overrun could be valuable to a project management team. While the method presented in this step was not used by Lee *et al* (2011) to define the goal of the application, the same goal could have been formulated with the presented method.

The method requires that an investigator, during any phase, but preferably within an early phase, reduce the uncertainty of the sustainable project success criteria of the construction project by using a specific type of data mining. If Lee *et al* (2011) formulated their investigation goal according to the method provided here, it could have been as follows: ‘To provide predictions of the cost overrun of a construction project, using a regression model, during the planning phase of the project’. This goal is specific and is clearly aimed at improving the management of a project by reducing the uncertainty within the planning phase of the cost of the project.

Goal Definition for Prediction of the Impact of an Accident Onsite: Following the method presented in this step the goal for this application could be set as, ‘Providing data-driven estimates, by use of a classification algorithm, of the time impact an on-site accident or safety issue will have on the timeline of a construction project.’ This goal is clear and would reduce the uncertainty during the initiation and planning phases regarding the duration of the project.

3.3 Data Acquisition:

A major strength of data and text mining is that it can be used to extract valuable patterns and information from data sources that are not usually considered. The knowledge gained from these sources can be used to make informed decisions or to provide reliable predictions. All data mining applications require the acquisition of data.

The data required is highly dependent on the goals of the applications. The investigator must understand the environment from which the data is being collected as well as the specific

application in order to collect the most appropriate data available (Aggarwal, 2015). In Case Study 6 by Nayak and Qiu (2005), the importance of documenting and tracking errors, faults, solutions, actions taken, project details, cost, time etc. is stressed as a prerequisite for data mining applications aimed at project management. The investigator must be cognisant of the different organisational roles within each project and the data each organisation might store.

As the data is collected for an investigation it is stored in the so-called *data warehouse* in its raw form (Zaki and Meira, 2014). The data can be collected from a variety of sources. Some valuable sources of existing data are:

- **Existing project databases and spreadsheets:** Large consultancies, government agencies, and private client bodies often have project databases where project information is stored.
- **Construction project documentation:** Construction projects produce a large amount of documentation throughout their lifecycles. Information can be extracted from these documents for data mining purposes (see Section 3.3.1).
- **Building information models (BIM):** The increased use of BIM and the ability to attach project documentation and other information opens the possibility to using building information models as sources of data (see Section 3.3.3.)

New data will need to be collected if the application requires data that does not yet exist or cannot be collected from existing sources. This can be time-consuming and expensive. Future data mining enterprises in any organisation will greatly benefit from a commitment to collecting and storing as much data as possible about every project in which it takes part.

The extraction of information from documentation, the documentation in a typical South African construction project and BIM as a source of data are discussed in the three following sections.

3.3.1 Information Extraction:

The overall aim of information extraction is to detect and obtain structured information from semi-structured or unstructured text. Once the information has been extracted from the text it can be presented to the end user or it can be used by computer systems such as databases or search algorithms to provide a specific service for the end user. Information extraction is a specialised sub-section of data mining. Some examples of information extraction applications are given below (Jiang, 2012):

- Information extraction is used in biomedical research to accurately retrieve relevant documents from the wider scientific community based on references to biomedical entities and to present it to the researchers. This precludes researchers having to read through enormous amounts of published literature in order to find possible links.

- The financial sector uses information extraction to seek out specific pieces of information from online news articles to facilitate day-to-day decision making. This enables them to generate lists of financial and world events that could impact the financial sector.
- Intelligence analysts search through large amounts of text for information regarding possible terrorist plots or other illegal activities.
- Internet search engines have moved beyond key-word or phrase searches and have moved into the realm of information extraction. This method achieves a better overview of documents and sites providing a more enriched representation of these documents to the user.

These examples illustrate the powerful, wide-ranging applications that smart, text-searching technologies can provide. Information extraction can be a useful part of data mining during the data acquisition phase and pre-processing phase.

Early information extraction systems used complex linguistic extraction methods that were developed by humans to match text and locate information (Sundheim, 1991; Chinchor, Hirschman and Lewis, 1993). These rule-based systems were able to achieve good performance in the target domain, but the rules are extremely domain specific and the definition of the rules is highly labour intensive.

Researchers realised the limitations of the manually developed rule-based systems and turned to statistical machine learning approaches. The decomposition of information extraction systems into subtasks such as ‘named entity recognition’ and ‘relation extraction’ transforms the nature of the problem into a classification problem, which further encouraged the use of standard supervised learning algorithms (Jiang, 2012):

- **Named entity recognition:** A named entity is a series of words that designates an information unit representing some real-world entity. The recognition and classification of names, places, times, dates and numeric percentages all falls under named entity recognition and forms an integral subtask of information extraction (Nadeau and Sekine, 2007). Named entity recognition and classification cannot simply be achieved by matching text against a pre-compiled list as the list would have to encompass all possible named entities, which is an infinite set, and any pre-compiled list would be incomplete. There are two approaches that have been historically used, the rule-based approach and the machine learning approach. The rule-based approach uses a set of rules, either manually defined or automatically learned to identify and classify named entities. The text is analysed using lexical analysis which assigns tokens to each word in the text. These tokens are compared against the rules and a rule is triggered if a match is found. A hierarchy of rules exist to handle cases where multiple rules are triggered (Jiang, 2012). The statistical machine learning approach is the second approach that has been used for named entity recognition. The machine learning approach treats the named entity recognition problem as a sequence labelling problem. Sequence labelling is a well-

established machine learning problem that has been successfully used to model many natural language processing tasks. Jiang (2012) provides an in-depth explanation of the implementation of the machine learning approach that will not be discussed here.

- **Relation extraction:** Relation extraction is the act of identifying and classifying the semantic relationships between entities mentioned in text. There are several different types of relationship that must be considered, examples of which include the physical proximity of two entities, the personal/social relationship between entities, and the employment/affiliation of two entities (Jiang, 2012). A variety of techniques has been proposed to successfully extract relationships between entities. The most common of these techniques is to approach the problem as a classification problem: given two entities occurring in a sentence, can the relation between the two entities be classified into one of the predefined relation types? The methods used to extract relationships between entities are feature-based classification, kernel based classification, and weakly supervised learning methods (Nadeau and Sekine, 2007; Jiang, 2012). These are all forms of supervised classification (see Section 3.5.5), that require relatively large labelled datasets.
- **Unsupervised machine learning:** The discussion of named entity recognition and relation extraction highlighted the need for well-defined entity and relation types in advance, based on the application, and having a large amount of correctly labelled training data as they are supervised machine learning applications. The creation and curation of these large training datasets and lists requires human expertise and a large amount of time. In an attempt to move completely away from these large training data sets, recent studies have looked at unsupervised information extraction from large corpora of semi-curated, freely available knowledge (Nadeau and Sekine, 2007; Jiang, 2012). These unsupervised information extraction methods include relation discovery, template induction and open information extraction which are all forms of unsupervised clustering (see Section 3.5.3). These methods aim to discover information entities and extract the most salient relationships between them in an unsupervised manner.

3.3.2 Typical Documentation in a South African Construction Project:

All construction projects produce a large amount of documentation that will be used at some point during the 5 phases of a project. The documentation is often stored by the party that creates and/or receives the documentation. Over the course of many projects, an organisation can accumulate a significant number of documents that contains large amounts of information about these projects. With the information extraction techniques discussed in the previous chapter, the possibility of extracting data from the documentation becomes feasible. The data extracted from these documents can be used to supplement data gathered from more conventional data sources.

In South Africa there are a number of common documents that are created during the life of a construction project. These documents can contain valuable information about the

management of the project that might not be available elsewhere. Figure 3-2 and the list below shows the relationships between clients, engineers, and construction contractors for a typical construction project and lists the documentation that is typically passed between the entities within the relationships.

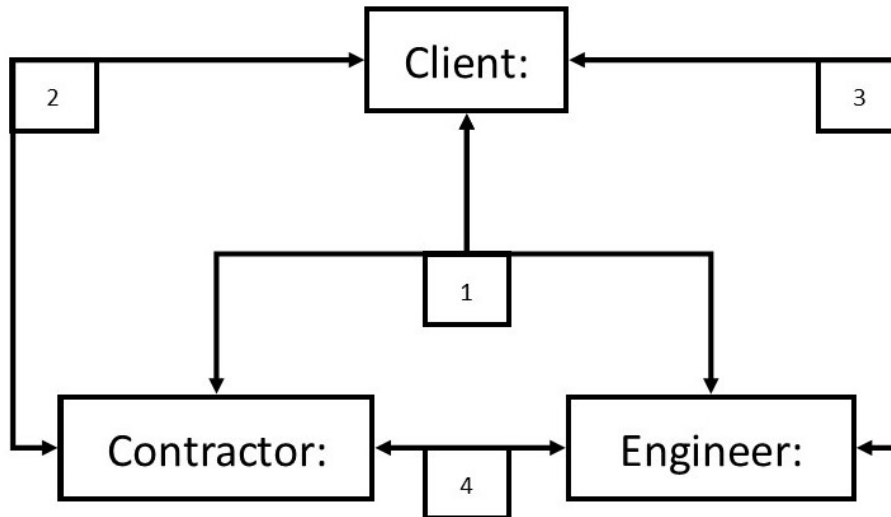


Figure 3-2: Client-Engineer-Contractor Relationships

1. The documentation that typically passes between the contractor, the engineer and the client are:
 - a. Bill of Quantities.
 - b. Payment Certificates.
 - c. Contract: Client – Contractor.
 - d. Change Order Letters.
 - e. Project Programme.
 - f. Project Cost.
 - g. Tender Documentation.
 - h. Email/Correspondence.
 - i. Environmental Approvals.
 - j. Lessons Learned.
2. The documentation that typically passes between the contractor and the client are:
 - a. Project Risk List.
3. The documentation that typically passes between the client and the engineer are:
 - a. Contract: Client – Engineers.
 - b. Statutory/Municipal Approvals.
4. The documentation that typically passes between the engineer and the contractor are:
 - a. Drawings: General/Layout.

- b. Drawings: Detail.
- c. Request for Information/Clarification.
- d. Responses to Requests for Information.
- e. Notifications of Delay.
- f. Site Meeting Minutes.
- g. Snag Lists.

The relationships shown in Figure 3-2 and the accompanying documentation are what can be expected from a typical construction project. Larger or more complex projects could have more stakeholders, resulting in more complex relationships and different documentation. There will also be idiosyncratic variations in the documentation and their depth depending on the particular organisations involved in the project. Nevertheless, typically the information contained in the documents above will need to be passed from one stakeholder to the other regardless of the format of the documentation.

The large pool of documentation that is created for each construction project is typically stored electronically after the project closure phase. The information contained in the documentation is a rich information source for data mining if the information can be extracted and added to the data warehouse by a practitioner that understands the environment and the value of the information.

3.3.3 Building Information Models as a Source of Data:

Building information models are data rich, three-dimensional, digital models of a building. The models are generated using specialised BIM software and a collaborative design process. The collaborative design process has been lauded as the most influential part of the BIM revolution as it allows design teams to collaborate on a single virtual model. This means that the different design teams present on a project can implement their designs and detect possible construction or usability issues in a wholly virtual context. The model also allows for additional information to be attached to the model such as information regarding cost, scheduling, structural performance, energy consumption and much more (Kensek, 2014).

The progress of BIM adoption has been quantified into the concept of BIM levels. Each BIM level represents a step towards the full virtual and collaborative design, management and maintenance possibilities of BIM. Each BIM maturity level is briefly described below (Construction Industry Council, 2013; NBS, 2014):

- **BIM Level 0:** BIM level 0 effectively means zero collaborative effort. Unmanaged 2D computer aided design (CAD) is used, mainly for production information. Drawings and other output are shared and distributed via paper, electronic prints, or a mixture of both but without common standards, formats and processes.
- **BIM Level 1:** BIM level 1 adoption means managed CAD, with increasing introduction of spatial coordination. This typically involves a mixture of 3D CAD for concept work, 2D for drafting, statutory approval documentation and production information. Standardised structures and formats are introduced and managed out of a

common data environment (CDE). The models are not shared between project stakeholders and team members. Level 1 is often described as ‘lonely BIM’ due to models not being shared.

- **BIM Level 2:** BIM level 2 is distinguished by collaborative work. The discipline-based models are created using managed 3D CAD with data attached. The parties use their own discipline-based 3D CAD models with data attached and do not work on a single, shared model. The collaboration comes in the form of how information is shared between parties. Design information is shared through common file formats which allows parties to incorporate other models into theirs to create a federated model and to carry out integrated checks on it. Data attached to the model may include construction sequencing (4D) and cost (5D) information.
- **BIM Level 3:** BIM level 3 is seen as the ultimate goal for the construction industry with all parties working on a single collaborative, online project model which is held in a centralised repository. All parties can access and modify the same model, effectively removing the last layer of risk for conflicting information. Software that specifies originator/read/write permissions are essential to reduce mutual liability and resolve copyright issues.

BIM has been used extensively in the pre-design and design phases of projects where large amounts of data about the project, such as possible costs, construction schedules, site surveys and more are stored. Increasingly, project documentation has also been attached to the model so that it is available to all the required stakeholders and is securely stored. BIM software developers aim to develop systems that will specifically handle this increasing demand and render the data easily accessible (Smith and Tardif, 2009).

The 4D modelling and tracking capability allows the BIM model to play a large role throughout the construction phase of a project. The construction site can be mapped out and the building virtually assembled to ensure that no clashes occur, or other problems arise. As the construction progresses, the project schedule can be updated with actual costs and durations. Better on-site visualisation and construction information reduces the risk of rework and other lack-of-information problems (Kiprotich, 2014).

The maintenance phase of a project can also benefit from BIM as detailed maintenance information of every system can be attached to the model. This enables a system manager to anticipate every maintenance activity and order required parts or services in advance to ensure optimal system operation and minimum downtime (Kiprotich, 2014).

The importance of BIM to the acquisition of data for data mining has not yet been fully explored or realised. BIM is used through every phase of a construction project and often contains large amounts of data about a project. This data is stored digitally and can be easily accessed by a project manager. These BIM models are stored for the entire operational lifetime of the project and thus could prove to be an excellent source of information and data for data mining. The project information, such as projected costs, projected durations, actual

costs, and actual durations about the construction activities can prove to be extremely useful for reducing uncertainty during the early stages of a new construction project.

The documentation attached to the model can also be a very useful source of information as it is all stored electronically in a limited number of formats. This reduces the difficulty of sourcing documentation and transforming it to a usable format and encourages information extraction to provide additional information to the data warehouse. A note must be made that while BIM is being used around the world, the application of BIM for medium and small projects in South Africa is still lagging. Data mining could be an added benefit to BIM adoption, if both are done correctly (Kiprotich, 2014).

3.3.4 Data Acquisition for Construction Project Management Examples

The discussion of the data mining examples defined in Section 3.1.3 is continued in this section. The goals for the applications were set in Section 3.2.2. This section discussed the acquisition of the required data for the examples.

Data Acquisition for Construction Project Cost Overrun Example: The data acquired for this application was collected for 77 construction projects in South Korea using a questionnaire and in-person interviews of experienced project personnel who worked on the planning of the selected projects (Lee *et al.*, 2011).

The questionnaire contained questions about actual cost as well as the initial duration and cost estimates from which the cost overrun could be calculated. A five-point Likert scale (from strongly disagree to strongly agree) was used in combination with the Project Definition Rating Index (PRDI – a project scope definition tool with 64 categories) to quantify the planning and scoping activities of each project (Lee *et al.*, 2011).

Data Acquisition for Prediction of Impact of Accident on a Construction Site: Construction sector workers are the most likely, compared to any other sector's workers, to experience an accident at work. These high accident statistics have led to procedures being adopted for recording and reporting of on-site accidents. The contents of these reports will vary depending on the legislative requirements of the country of operation but typically contains information regarding the nature of the accident, the severity of the incident and the number of workers injured (Tixier *et al.*, 2016).

Information extraction methods may be used to obtain information regarding the nature and severity of the incident from accident report documentation. If stored digitally, the documents may be appropriate for text mining (see Section 3.5.6). Project information such as the initial and final duration and cost estimates should be collected. If possible, data regarding the scope/work undertaken during the project should be collected. Project monitoring information such as project Gantt charts may also be collected. If information about the impact of the accident is directly available that data should be collected. This data could be found on internal project databases or within internal health and safety reports.

3.4 Pre-processing:

The pre-processing of the data is an important part of the data mining process that requires the carefully thought out application of several data handling techniques to ensure that the data is adequately prepared for use in a data mining application (Aggarwal, 2015). The goal of the pre-processing step is to produce a clean, informative, and normalised target dataset that is suitable for the selected data mining model type. The data from the data warehouse can be widely variable in both format and quality. Missing values, errors and inconsistencies are some of the data problems that must be resolved to produce a good target dataset.

The first sub step within the pre-processing step is the re-examination and possible re-definition of the goals of the application. The goals of the application are re-examined to determine if they are feasible using the available data. Once the goal of the application has been re-examined, the target dataset can be assembled. The second sub step in the pre-processing step contains three different data processing techniques to assemble a clean, informative target dataset. The data processing techniques are:

- **Feature Extraction and Selection:** The most informative features are extracted from the data warehouse, if necessary, and selected to produce a set of features that can accurately model the system.
- **Data Cleaning:** Missing values, data entry errors, inconsistencies and scaling of the data are dealt with.
- **Data Reduction and Transformation:** Excess or uninformative data is removed from the target dataset. The data is transformed into more informative features or features of a different data type.

The typical process for applying these data pre-processing techniques to a dataset is shown in Figure 3-3 (Kotsiantis, Kanellopoulos and Pintelas, 2006; Srivastava, 2014):

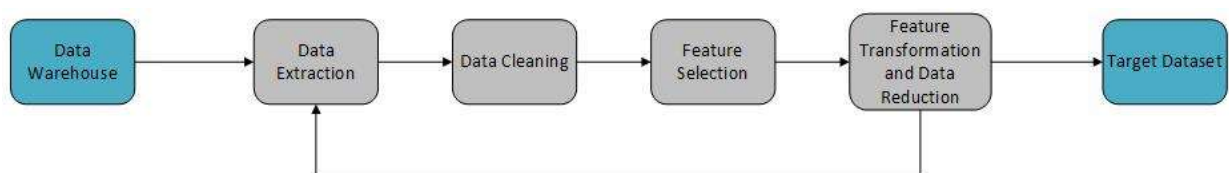


Figure 3-3: Data Pre-processing

The two sub steps and four data processing techniques will result in a target dataset that is well-suited for the goal of the application, informative, appropriate for the selected model, and free from ‘noise’.

3.4.1 Goal Re-examination:

Once the data has been collected and the investigator has had time to examine and become familiar with the available data, the goals of the investigation need to be re-examined. The goals of the investigation were initially set to inform the data acquisition process and will influence every other phase of the data mining process. The goals of the application are to provide information to the project team to reduce the uncertainty within a project phase regarding a success criterion (Cost, Time, Quality, Social Impact, or Environmental Impact). At this point the goals should be re-examined to determine if the data mining application is possible with the data that has been collected. This can be achieved by a preliminary investigation of the data. Examining the data for strong correlations will allow the investigator to determine which features might be valuable, which features should be ignored, if there are many outliers, and other possible problems with the data. The goals or scope of the application might need adjusting if little or no correlation exists within the data.

If additional data must be collected it is up to the investigator to decide how to proceed. If the scope of the application needs to be altered or be expanded, the investigator must determine the extent of change and what the implication of the changes will be. An example of this was the narrowing of scope by Fan, Abourizk and Kim (2008) in Case Study 1 to determine the residual value of only one type of heavy construction equipment.

The iterative nature of the data mining process allows the goals of the applications to be revisited several times throughout the implementation of the data mining application should unforeseen issues arise that affect the viability or scope of the application.

3.4.2 Feature Extraction and Selection:

The data warehouse contains vast amounts of raw data in the form of text documents, project information stored in databases, payment and transaction information, legal documents, tender information and much more. It is the analyst's task to examine all the data and to select the data fields (referred to as features, attributes or dimensions) that will be used in the investigation. Depending on the contents of the data warehouse, the investigator might have to apply text mining (see Section 3.5.6) or information extraction to text documents to obtain the desired features from the raw data.

If all possible features of a system can be determined and recorded with perfect accuracy, then the system could be modelled perfectly. However, this is impossible for nearly all applications. The data to perfectly model a system does not exist and data mining models typically do not work effectively with very high dimensional data (Aggarwal, 2015). The investigator should therefore select the features that are most relevant to the application.

Features that are best suited for inclusion in the target dataset are the features that will influence the outcome of the data mining application. For example, the height of an individual is not a feature that will influence whether the person will donate to a charity organisation. The income of a person is much more likely to influence whether a person

will donate to a charity and should therefore be selected as a feature to include in the target dataset.

Features that have internal relationships, such as an individual's address and their postal code, should not both be included in the target dataset. This precludes the introduction of redundancies and noise arising from the same or similar information. There are a variety of techniques to programmatically identify and remove irrelevant or uninformative features from the dataset, such as sampling, feature subset selection and dimensionality reduction. Other methods such as Wrapper models and Filter models algorithmically determine the most appropriate features to include in the target dataset (Han, Kamber and Pei, 2012; Zaki and Meira, 2014). The specific requirements for selecting features are discussed in Section 3.5 Mining and Modelling for each mining model type.

3.4.3 Data Cleaning:

The data cleaning process is the step in the pre-processing phase where many of the issues with the data will be dealt with. To produce accurate and reliable results, data mining and machine learning models need accurate and reliable data. During the collection process there are several sources of inaccurate data, missing entries or errors in the data (Aggarwal, 2015):

- Data collection hardware, such as sensors, may be inaccurate or un-calibrated.
- Scanning and character recognition systems can make mistakes when scanning documents or other files.
- User-entered data is likely to contain inaccuracies, mistakes, missing values and errors since manually created data is more prone to error.
- The entity that oversees the initial data collection may not value certain types of data or think that collection thereof is unnecessary or too costly. This could result in lack of detail or large amounts of missing data.

If the issues mentioned above are not properly dealt with they will be a significant source of inaccuracy in the data mining application. There are three basic aspects of data cleaning (Han, Kamber and Pei, 2012; Aggarwal, 2015):

1. **Handling missing values:** Missing values can be handled by deleting the entire record with the missing value. This is not always acceptable, such as when there is limited data available or the data has a high missing value count. In these cases, the missing data can be estimated or imputed. The imputation process itself can be a large undertaking but generally with dependency-orientated data, such as spatial or temporal data, contextually nearby values can be used to estimate the missing value. For example, the mean of the values before and after a missing time series value can be used to estimate a value. In non-dependency orientated data, a classifier (see Section 3.5.5) can be used to estimate the value.
2. **Handling incorrect and inconsistent entries:** The combination of domain knowledge, inconsistency detection and outlier removal are methods used to handle incorrect and inconsistent entries. Domain knowledge can be used to set limits on

acceptable values for attributes and can be incorporated into data auditing tools that are common in database management. These tools allow the investigator to detect inconsistencies. For example, if the country entry is ‘Germany’ the city entry cannot be ‘Shanghai’ or ‘Italy’. Outlier removal can be used in data-centric models. Outliers can be detected by using basic statistical analysis or by using more sophisticated outlier analysis methods (see Section 3.5.4)

3. **Scaling and normalisation:** The scale of the collected data fields can vary dramatically due to the nature of the data. For example, the number of sub-contractors on a construction project and the cost of the project have vastly different scales. If the scales of data fields are skewed, the similarity computation (see Section 3.5.1), will be skewed to regard the data of larger scale to be more influential. Therefore, data must be scaled or normalised to ensure that the different data fields are examined on an even footing.

Carefully examining and thoroughly cleaning the data is a worthwhile step as it will help the investigator create a clean, representative target dataset for effective data mining.

3.4.4 Data Reduction and Feature Transformation:

Data mining algorithms can be computationally expensive. The target dataset should therefore be as compact and representative as possible to ensure quick training times and model usability. The data can be reduced by decreasing the number of columns (features/attributes/dimensions), or by decreasing the number of rows (data entries/records). This process does cause data to be lost but can provide great gains in computation speed and usability. Smaller datasets allow for the use of more complicated and sophisticated data mining algorithms. Therefore, the dataset reduction will not automatically lead to less accurate results. There are four data reduction methods that can be used (Han, Kamber and Pei, 2012; Aggarwal, 2015):

1. **Data sampling:** By taking a strategic sample of the dataset to compile the target dataset, the entire dataset can be represented by a much smaller dataset. This is typically done by randomly selecting a predefined fraction of the total dataset without replacement. By selecting without replacement, the subset will provide an accurate representation of the entire dataset as no data entry will be repeated. The selection process could be pseudo-random in that it selects a fraction of the dataset while ensuring that the full range of values are present in the target model. This is known as biased sampling, which is useful in reducing the size of datasets used for outlier detection while still maintaining the outliers within the data. Another example of biased sampling is sampling within some predefined and desired ‘stratum’ in the dataset. This is typically used if the dataset does not contain enough entries of a certain ‘stratum’ and the investigator wishes to ensure that each ‘stratum’ is equally represented in the target dataset.
2. **Feature subset selection:** This section overlaps with feature extraction and selection (see Section 3.4.2). The dataset may contain features that are irrelevant to the

application and can be discarded. The features that can be discarded are extremely application and domain specific and can rely very heavily on the judgement of the investigator and their knowledge of the domain.

There are some unsupervised and supervised machine learning approaches that will help with features subset selection. These advanced methods are based on clustering and classification (see Section 3.5.3 and 3.5.5 respectively).

3. **Dimensionality reduction with axis rotation:** An entire dataset of truly non-dependency orientated data is typically quite rare. Most datasets contain data that has some internal relationships, either explicit or implicit. These relationships can be used to express multiple dimensions as one dimension. Automatic methods such as principle component analysis and singular value decomposition can be used to represent a number of dimensions as one dimension. An example of this is the singular value decomposition done by Williams and Gong (2014) to reduce a sparse bag-of-words text matrix to a single dimension to be used in a data mining algorithm. They achieved greater accuracy when the decomposed text was present in the target dataset.
4. **Dimensionality reduction with type transformation:** Some data can be difficult to integrate into a target dataset for a data mining model. In the case where the data is of a complex type, the data can be reduced and transformed to a simpler type using dimensionality reduction and type transformation. For example, data from a time-series cannot easily be combined with other data from a dataset and can be transformed by using binned values or more complex methods such as Haar wavelet transformation.

Depending on the data mining application and domain, the investigator may be able to perform the data transformation step without the assistance of sophisticated methods. For example, the start and end date of a construction project may be simply transformed into a duration. This type of intuitive transformation is a good starting point for novice data mining practitioners when conducting data reduction and transformation. If the dataset is very large or relatively complicated, the more sophisticated methods can be used to reduce and transform the data.

Reducing and transforming data provides more information to the target dataset and as such is a valuable tool in data pre-processing due to the resulting improvement in computation efficiency and algorithm accuracy. Once the goal of the investigation has been re-examined, the useful features have been extracted from their original forms in the data warehouse, the data has been cleaned, the data has been reduced and transformed the target dataset should be ready for the data mining model selected for the application.

3.4.5 Pre-processing for Construction Project Management

Examples:

The discussion of the data mining examples defined in Section 3.1.3 is continued in this section. The goals of the applications were set in Section 3.2.2 and the data acquisition was conducted in Section 3.3.4. This section discusses the pre-processing of the data.

Pre-process for the Cost Overrun Example: Once the interviews had been conducted and the project scope and planning information from the Project Definition Rating Index (PDRI) was collected the goal of the application should have been reviewed. While Lee *et al* (2011) did not explicitly follow this step, they did conduct an initial investigation into the collected data to determine any data issues.

They found that not all the PDRI questions were applicable to every project, resulting in 3.25% of all values missing from the questionnaires. Several inconsistencies in the scale of the base cost of the data was also noted. These data quality problems were not severe and the investigators continued with their application (Lee *et al.*, 2011).

The missing values were imputed using a k -nearest-neighbour algorithm (see Section 3.5.5.5 for classification configuration) that assigned the mean value of several data points that were similar to the data point with the missing value. The scale inconsistencies were rectified, and the initial cost estimate and final cost were transformed into a cost overrun feature (in %) (Lee *et al.*, 2011)

The feature selection was conducted using a wrapper feature selection method (see Section 3.5.5 for a brief discussion of wrapper methods). Once the wrapper feature selection method had selected the most informative and appropriate features for a regression algorithm, the target dataset was assembled from the clean data (Lee *et al.*, 2011).

Data Pre-processing for On-site Accident Example: The data collected for this investigation should cover two aspects: 1) The project information (such as the project duration and cost), the scope of the project and the project monitoring information; 2) The accident information from the accident reports. The goal of this application was set as the estimation of the impact on the duration of the project of an on-site accident by use of a classification algorithm. The model could be used during the risk identification and analysis steps of project management. This will allow the project management team to more accurately estimate the impact of safety risks.

The information collected should cover the type of work that being done on the site, precursors to the accident, and the details of the accident and the impact of the accident. The data should be investigated for defects such as outliers, inconsistencies and missing values. The data collected should then cover all the information required to model the problem. If the data does not have a large number of problems, then the application can continue.

Any data quality issues such as missing values, inconsistencies, large scale differences and outliers must be dealt with. This can be done by removing the offending data entries or

altering them with the appropriate tools (normalisation, imputation etc.). The initial duration and the final duration should be transformed to the % time overrun along with any other necessary feature transformations.

Since the model that will be used is a classification model, the time overrun (%) should be binned into 5 categories. These 5 categories will correspond to the ‘Low Impact’ to ‘High Impact’ categories commonly used in a project risk register. The investigators will need to define what they consider as ‘Low Impact’ and as ‘High Impact’ and bin the projects into the appropriate categories. These risk impact categories will be the class assigned to each project and thus the output from the model. The features of the dataset should be selected using a filter model. The most informative features will then be included in the target dataset.

3.5 Mining and Modelling:

The choice of which data mining model and data mining techniques to use along with appropriate data validation techniques is an important part of data mining. The selection of the appropriate model is governed by many factors such as the type of data, the quantity of data, whether the data is labelled, and the desired output from the model. The large amount of possible data mining and machine learning techniques and applications make it a complex and sometimes overwhelming environment to enter if the investigator is not already well-versed in statistics and machine-learning. The following section aims to provide a basic explanation and summary of several different types of data mining. Much of the information in this section is drawn from Aggarwal (2015), Zaki and Meira (2014), and Han, Kamber and Pei (2012). The following topics are discussed:

- **Similarity and distance functions:** determining the similarity between data points. This forms the basis of the majority of data mining models.
- **Association pattern mining:** mining patterns from data based on the probability of data points occurring together. This is widely used in supermarket shopper analysis.
- **Cluster analysis:** using various similarity and distance functions to determine if data points belong together in groups.
- **Outlier analysis:** using similarity and distance functions to determine which data points do not belong in a cluster or do not follow the patterns within the data.
- **Classification and regression:** using examples to extract patterns from the data to predict the class or value of unlabelled data points.
- **Text mining:** using specialised techniques to represent text as multidimensional data that can be used in other models, such as classification.

This section is not an exhaustive list of all possible data mining types and algorithms but rather serves to prime a new data mining practitioner with the many possibilities. It is rarely

the case that a specific data-mining application will use all the methods described below. In fact, a data mining application will typically only use a few of the following methods.

3.5.1 Similarity and Distances:

Data mining techniques, models, and applications typically require the similarity or dissimilarity of objects, attributes, events and patterns within the data to be determined. A method for quantifying the similarity of data objects is required for data clustering, classification and regression, outlier detection, and virtually all other data mining applications. Similarity and distance are two inverse representations of the same concept. The greater the similarity between two data points, the smaller the distance between those two data points. The context of the data can be significant. For example, when using spatial data, it is more intuitive to use a distance measure, but when using a matrix representation of a piece of text it is more intuitive to use a similarity measure (Aggarwal, 2015).

Distance functions are a fundamental part of an effective data mining algorithm as almost all data mining algorithms perform their function by first computing how similar the data points are. A poor or inappropriate distance function will have a detrimental effect on the accuracy of the data mining model. Whether a distance (or similarity) function is appropriate will be determined by a) the data types, b) the data distribution and c) the data dimensionality. Similarity and distance functions are highly sensitive to these factors and they must be taken into account when selecting a specific function. Of the three factors, data type is the largest influencing factor. The functions are therefore generally split up into the main data types (Aggarwal, 2015):

- **Multidimensional data:** this type is the most common form of data but there exists a large amount of diversity within the data type, such as categorical or quantitative data. Within this data type several different methods are designed to function under various data distribution and data dimensionality constraints.
- **Text data:** text data can be seen as a special form of multidimensional data if the bag-of-words approach is used, as is most often the case (see Section 3.5.7). The very sparse and non-negative nature of this representation of text data warrants several methods dedicated to accurately determine the similarity between different pieces of text.
- **Temporal data:** the two-part form (contextual - time, behavioural – recorded value) of temporal data creates several unique challenges to determining the similarity of the data.
- **Graph data:** The similarity functions for graph data can be split into those that determine the similarity between two sets of graph data and those that determine the similarity between two nodes in a single graph. Methods vary largely in their implementation and how they determine the similarity of the graph data.

While the similarity and distance functions form the bedrock of many data mining models, only very sophisticated data mining applications will have a custom user-defined similarity function. All the predefined data models that exist have their similarity function built in. The exact implementation of these functions are therefore outside the scope of this research and will not be discussed further. However, if a data mining practitioner wishes to create their own, or to customise an existing distance function, it is recommended that they consult one of the three textbooks presented in Section 3.5.1.

3.5.2 Association Pattern Mining:

Association rule and pattern mining is an unsupervised learning technique that is used to identify relationships between different attributes within the dataset. The technique was first defined in the context of shopper analysis for supermarkets. The method defines the level of association of a set of attributes in terms of the frequency of occurrence. This is why this task is also known as Frequent-Itemset Pattern Mining. Apart from this initial use it has seen use in text mining and other dependency-oriented data (Srivastava, 2014). An association rule is typically shown in the following form:

$$\{A\} \Rightarrow \{B\} (\text{Conf: } i\%; \text{ Sup: } n)$$

Where $\{A\}$ and $\{B\}$ are a set of attributes, commonly called an item list. *Conf i%* refers to the conditional probability that $\{A\} \cup \{B\}$ occurs in a data record given that the record contains $\{A\}$. *Sup: n* refers to the support for the rule, or the frequency of occurrence of the rule (Aggarwal, 2015). An example of an association rule is:

$$\{\text{Eggs, Milk}\} \Rightarrow \{\text{Yoghurt}\}$$

This rule states that when someone buys eggs and milk, they are also likely to buy yoghurt. The usefulness of these types of rules is obvious as they are easily interpretable by a human and can easily be acted upon. The manager of the shop can then place the yoghurt between the milk and the eggs in their shop to promote the sale of yoghurt. Association pattern mining typically has two stages (Zaki and Meira, 2014):

- **Phase 1:** Generate all association rules with a support level above a user defined minimum (*minsup*). The generated rules must be supported by a sufficient number of examples in the dataset to be accepted.
- **Phase 2:** Trim all the generated association rules with a confidence level below a user defined minimum (*minconf*). The rules that do not have sufficient strength, in terms of the confidence of the conditional probability of the two item lists occurring together, are removed.

Association rule generation is a computationally intensive exercise for large datasets or datasets with high dimensionality. The trimming phase of the rule formulation is far more straightforward, as each rule's confidence is checked against the user specified *minconf* and is discarded if it does not meet it. The majority of the research in this area is focused on the first

phase, where the rules are generated. Several methods based on the frequency of occurrence have been devised for generating association rules such as the Apriori algorithm, Enumeration Tree algorithms, and Recursive Suffix-based pattern growth methods that aim to reduce the computational intensity of generating the rules while extracting all the noteworthy rules (Srivastava, 2014; Aggarwal, 2015).

Alternative models do exist that do not base their level of association on the frequency of occurrence of the item lists in the data. These algorithms attempt to solve the issue that raw frequency of item list occurrence does not necessarily result in interesting patterns being discovered. These alternative algorithms attempt to use more robust statistical methods to determine application specific rules that are interesting but suffer from several issues related to proper algorithm generation. The vast majority of association pattern generation algorithms in use are therefore still those traditionally based on frequency of occurrence (Srivastava, 2014; Zaki and Meira, 2014; Aggarwal, 2015).

Other possible solutions to the uninteresting rule problem are the number of summarising and querying methods that have been created to examine the generated rules and to abstract knowledge from them. These techniques allow more generalised rules to be synthesised from the base set of generated rules and allow the investigator to easily determine which are noteworthy and which are not. The implementation of these methods is discussed by Aggarwal (2015) and is beyond the scope of this investigation. Association pattern and rule-making algorithms are often incorporated as a subroutine in other data mining applications such as clustering, classification and outlier analysis, which are discussed later in this chapter.

3.5.3 Cluster Analysis:

Clustering is the unsupervised learning process of analysing a given set of data and partitioning the data into smaller groups, or clusters, containing data points that are very similar (Alsabti, Ranka and Singh, 1997; Aggarwal, 2015). At its core, clustering can be considered as a type of data summarisation. Clustering is often used as a sub-routine in larger data mining applications as the data summarisation helps to extract concise insights from the data. The clustering problem can be formulated in a large variety of ways and is influenced by such factors as the objective criteria for similarity of the data points and the number of clusters that are present in the data (Han, Kamber and Pei, 2012).

At its base level, clustering is the process where an algorithm determines the relative similarity between data points and, based on their similarity to each other, assigns them to a cluster. The difference between most clustering models is that they use different similarity measures and cut-off points for determining if a data point belongs to a cluster. Some clustering algorithms determine the relative similarity between data points with a distance measure (such as Euclidian distance), while others use density-based models or probabilistic-mixture models (Witten and Frank, 2005).

Regardless of the base similarity function, the selection of influential features and removal of features that cause noise or other distortion is critical for a successful clustering application.

Feature selection for unsupervised learning processes such as clustering or outlier analysis tends to be a more difficult task than feature selection for supervised processes such as classification and regression. This is because supervised learning processes have the correct label for the data point available to them. The features can then be examined for their ability to predict the class to which a data point belongs. This cannot be done for unsupervised learning processes (Witten and Frank, 2005; Zaki and Meira, 2014).

Two primary methods exist for conducting automated feature selection: filter models and wrapper models. These models use different methods to examine a set of features to determine their inherent clustering tendency. Filter models use predefined criteria to determine the clustering tendency of certain features and provide a numerical score for each feature's clustering tendency. Wrapper models use a clustering model along with internal cluster validation criteria (see Section 3.6.1) to conduct an iterative investigation of the clustering tendency of the features. The wrapper model forms part of the overall clustering process and, unlike filter models, cannot be conducted separately (Witten and Frank, 2005).

Once the features have been selected to reduce noise, the clustering algorithms conduct the analysis on the target dataset. The algorithms typically require the investigator to specify a few parameters, such as the number of clusters, the density, or the rank of matrix factorisation. These parameters will depend on the specific clustering algorithm selected. Understanding how these parameters affect the algorithm is important for ensuring optimal application of the clustering algorithm (Alsabti, Ranka and Singh, 1997; Han, Kamber and Pei, 2012).

Figure 3-4 shows a two-feature example of clustering and outlier analysis. Three clusters were identified along with several outliers. The bounds of the cluster are hard to identify as there are data points that could belong to a cluster or could be an outlier. The algorithm used assigned the data points to the clusters as shown but a different algorithm might have expanded or reduced the bounds of the clusters depending on the underlying similarity function. These fringe cases pose the greatest challenge for clustering and outlier analysis.

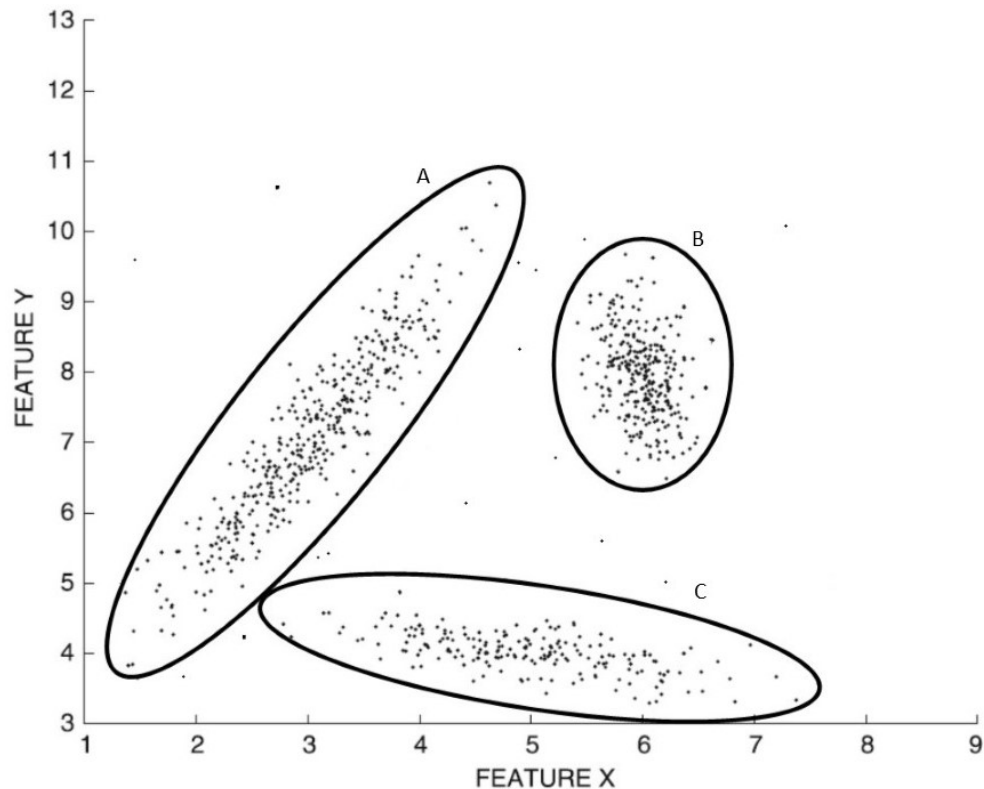


Figure 3-4: Clustering and Outlier Detection Example.

3.5.4 Outlier Analysis:

Outlier detection is a complementary process to cluster analysis. Where cluster analysis is concerned with trying to divide the data into clusters of similar data, outlier detection is concerned with detecting individual data points that differ from the rest of the dataset (Hawkins *et al.*, 2002). Both outlier analysis and clustering are unsupervised machine learning applications. Outliers are also referred to as ‘discordants’, ‘deviants’, ‘anomalies’, or ‘abnormalities’ in the data mining literature. Outlier detection can be used for two very different applications:

- **Data cleaning:** outlier detection can be used to detect and remove errors and noise in the data that may have arisen from the data collection process or some other source of noise (see Section 3.4.3)
- **Detecting and analysing rare events:** outlier detection forms part of the data mining process that generates information about a dataset. For example, outlier detection can be used to detect situations where something has gone horribly wrong, or incredibly well. In these cases, the investigators can examine these outliers and detect what happened. Other examples where outlier detection is very useful are credit card fraud detection and network intrusion detection, where unusual transactions or network traffic is detected and flagged, usually saving someone from fraud or a network from invaders.

Outlier detection works by examining the data and creating a model of the normal patterns within the data. This is typically done by using dimensionality reduction, clustering, or distance-based quantification. Once a model has been created that can mimic the underlying patterns within the data, the data is examined for data points that do not match these underlying patterns. The outliers are defined either by assigning a binary label or a real-value outlier score that determines their “outlier-ness”. Most algorithms typically provide a real-value score to quantify how much of an outlier a specific data point is. To arrive at a binary label, the algorithm chooses, or is provided by the investigator, with a cut-off point that determines which values are outliers and which values are not. The requirements for the features that are selected for outlier analysis are nearly identical to the requirements for the features used in clustering. This is because clustering and outlier detection are complementary processes. The clustering features selection process can therefore be used for outlier analysis (Hawkins *et al.*, 2002; Zaki and Meira, 2014).

A two-feature cluster and outlier analysis is shown in Figure 3-4. Three clusters were identified along with several outliers. As can be seen, there are several data points that fall outside the bounds of the identified clusters. These outliers are difficult to identify because different clustering algorithms could include certain data points in clusters whereas others might exclude these marginal cases. These marginal cases are the most difficult to ‘correctly’ cluster or label as an outlier. Despite these challenges, outlier detection can still provide valuable insight into a dataset.

3.5.5 Data Classification and Regression:

Clustering and outlier analysis algorithms examine an unlabelled dataset to learn the underlying data patterns to partition the data into data points that are similar and to identify data points that do not conform to these underlying patterns, respectively. Classification and regression are the related processes where an algorithm trains a model on a labelled dataset by determining the underlying patterns between the data features and the data label. Based on the learned patterns within the data, a classification model assigns a categorical or discrete label to an unlabelled dataset whereas a regression model assigns a continuous numerical label to an unlabelled dataset (Witten and Frank, 2005; Srivastava, 2014).

Classifiers and regression models are supervised machine learning applications since an example dataset is used to train the algorithm. This contrasts with the unsupervised learning approach of cluster analysis, where no labelled training dataset is used. Classification and regression models are typically trained on a subset of the labelled dataset and then tested on the remaining subset of the labelled dataset. Data classification and regression are more powerful and directly usable than data clustering as classification and regression uses external user-defined grouping from the example dataset. The groups are therefore already well defined and meaningful to the investigator. The advantage of usability has led to classification and regression being used in a large variety of applications: (Aggarwal, 2015):

- Customer target marketing: The classifier is given a dataset containing demographic profile features (age, income, gender etc) and a label of buyer or non-buyer,

corresponding to the buying history of the people and whether they have bought a specific product in the past. The classifier is trained on the dataset and learns the underlying patterns that exist in the dataset that can be used to predict whether a person, with no buying history label, is likely to buy or not buy the product.

- **Medical disease management:** Patient diagnosis, biometric data and medical tests are contained in a dataset along with the label of a specific treatment outcomes. The model will then be able to predict which patients are likely to respond well to certain treatments and which are not.
- **Document categorisation and filtering:** Web portals require the real time classification of large volumes of documents into categories such as sport, politics, current events etc. A large number of these types of documents have already been sorted into these categories. The classifier uses the words contained in the documents as features and the category and the class label. The classifier will then be able to automatically classify these documents into their categories.
- **Multimedia data analysis:** Photos, videos and audio files can be classified into sub-categories using classifiers. Photos can be classified into those that contain cats, dogs and humans. Music files can be classified into their genres. These types of advanced classification often use multiple machine learning models working in conjunction to classify the inputs, but the basic process of training on a labelled dataset, testing on another labelled dataset, then being deployed on unlabelled data remains the same.

The ability of classification and regression models to learn by example opens their application possibilities. As long as the investigator can formulate their problem in a suitable manner, the data can be modelled, the underlying patterns learned, and these patterns can be applied to other datasets.

As with outlier analysis, the output from a classification algorithm can be either a label or a numerical score that shows the likelihood of the data point belonging to a class. The model either automatically learns what an appropriate cut-off point is, or the investigator can specify the desired cut-off point. For a regression algorithm the output is just a numerical value that corresponds to the continuous label, much like conventional regression.

As with all data mining applications, the selection of features that are relevant to the problem that is being modelled is of vital importance. Though irrelevant features will typically have only a slight negative impact on the accuracy of a model they will be a significant source of computational inefficiency (Aggarwal, 2015). For example, the skin colour of a person is not an important feature to consider when predicting a ‘diabetes’ disease label whereas the age of a person is a far more important feature. There are three primary method types, beyond manual selection, that are used for feature selection for classification algorithms (Witten and Frank, 2005; Han, Kamber and Pei, 2012):

1. **Filter models:** each feature in a dataset is examined for its ability to predict the output class of a data point. A numerical score is assigned to each feature and the

features that fall below a given threshold are removed from the dataset. This is done independently of the classification or regression model.

2. **Wrapper models:** wrapper models form part of the classification or regression process and can tailor the features to be the most informative features for the specific classification algorithm.
3. **Embedded models:** these models are embedded into the classifier and can recursively train a classifier on the data and then eliminate the least valuable feature. This is done until no more gains in accuracy and efficiency are made, after which the final classifier is trained on the remaining features.

In the following sections, some of the commonly used classification methods will be briefly examined in order to familiarise the reader with the algorithms that are most often used and are widely available in predefined data mining resources (see in Chapter 4). While the explanations apply to classification algorithms, almost all the algorithms mentioned are able to perform regression modelling with some small changes. The internal functioning of classification and regression models is similar, and it is therefore not necessary to explain exactly how regression is performed.

3.5.5.1 Decision Trees:

Decision trees have been used for a wide variety of applications, but their application as a classification methodology has seen a large degree of success. The hierarchical decision system is used as a classifier system by making decisions based on feature variables. Each node of the decision tree has a split criterion, which is based on one or more features of the training dataset. The decision splits the dataset into two or more subsets (Witten and Frank, 2005).

Take, for example, the split criterion of $Age > 50$. In this case the 1st branch off the node will contain all the data in the training set where the *Age* feature is greater than 50 whereas the 2nd branch contains all the data in the training set where the *Age* feature is smaller than 50. This can then be further broken down with a sub-node on each branch with the split criteria of $salary > \$60\,000$ and $salary > \$50\,000$. This decision tree is shown below in Figure 3-5 and models whether a person is a donor to an organisation. The split criteria for any node may contain multiple features and multiple branches stemming from it.

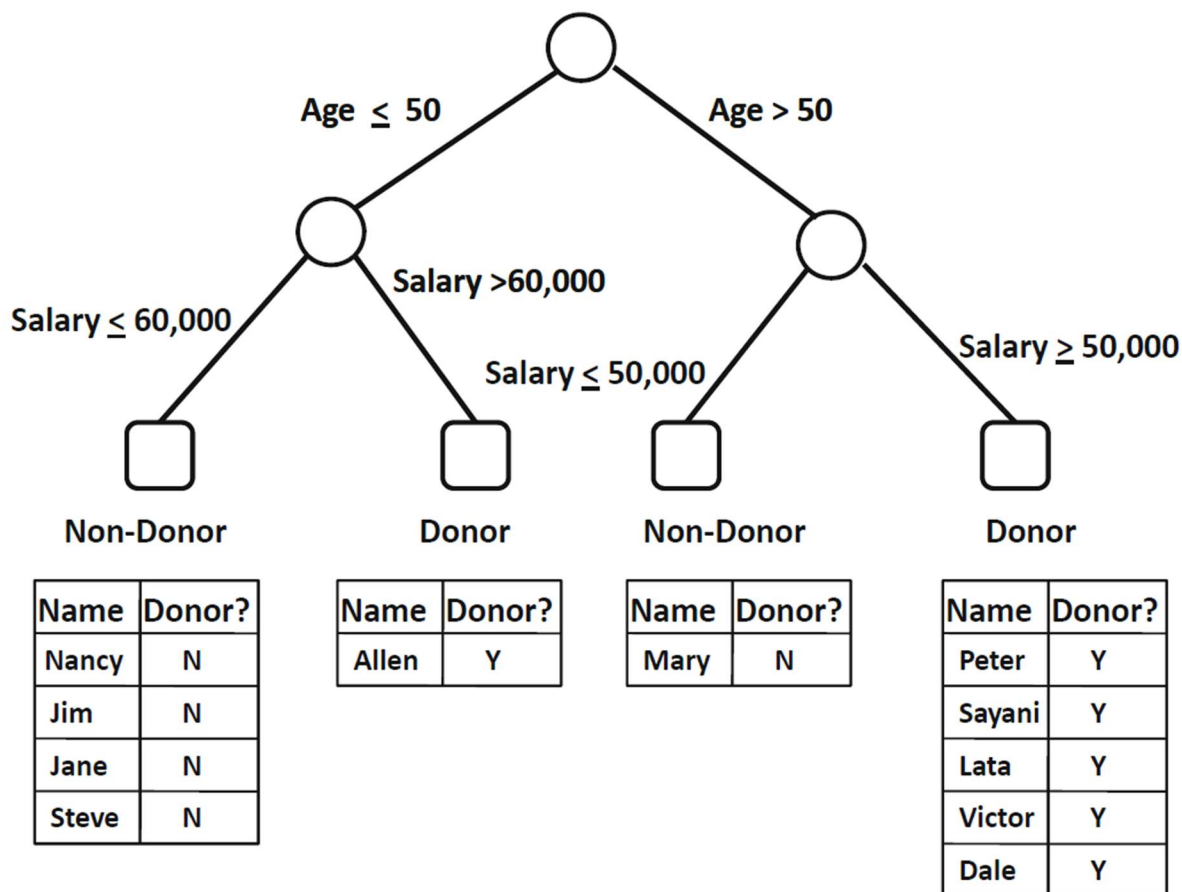


Figure 3-5: Decision Tree Donor Example (Aggarwal, 2015).

The aim is to find the smallest set of decisions that can model the dataset accurately. The number of decisions and branches can be trimmed in an iterative process to achieve an optimal decision tree (Zaki and Meira, 2014). One of the advantages of a decision tree is that the rules can be extracted and easily interpreted by the investigators, as was done by Nayak and Qiu (2005) and Lee, Hsueh and Tseng (2008).

3.5.5.2 Rule-Based Classifiers:

Rule-based classifiers generate a set of “If-Then” rules of the form:

$$IF \text{ Condition } THEN \text{ Conclusion}$$

The ‘condition’ and ‘conclusion’ are also known as the antecedent and consequent respectively. The condition may contain a logical operator, such as: $>$, $<$, \leq , \geq , $=$, \neq or \in . The logical operator is applied to the feature variables in the condition, while the conclusion contains a class variable (Aggarwal, 2015). Another form the rule-base classification can take is similar to the association rules form (see Section 3.5.2):

$$IF \{A\} \Rightarrow \{B\}$$

Where \Rightarrow corresponds to the THEN in the original form. An example of a rule from a rule-based classifier is shown below:

$$IF \{Age \leq 50 \text{ AND } Salary \geq \$60\,000\} \Rightarrow \{\text{Donor}\}$$

This rule is effectively the same rule as the combination of the left branch and subsequent right-sub branch of the decision tree in Figure 3-5. Rule-based classifiers often use the same rule generation methods as decision trees, but they can combine many conditions into one rule. This can also be done by decisions trees but is less common.

Rules are generated that are either mutually exclusive or exhaustive. Mutually exclusive rules cover a subset of the data, thus only one rule can be triggered by a test case if the rules are mutually exclusive. Exhaustive rules can cover the entire dataset and can overlap; therefore, more than one rule can be triggered by a test case. Since a test case can only be ascribed one class, exhaustive rules are either ordered into a priority list that determines what class a test case will be assigned or alternatively, the dominant class label amongst all the triggered rules is assigned to the test case (Witten and Frank, 2005).

3.5.5.3 Probabilistic Classifiers:

Probabilistic classifiers model the relationship between the feature variables and the class variable and quantify the relationship as a probability. Two of the most common probabilistic classifiers are the Naïve-Bayes classifier and the Logistic Regression classifier, but many more exist. The Naïve-Bayes algorithms use Bayes theorem to quantify the conditional probability of a feature set having a certain class label based on the prior conditional probability of the feature sets contained in that class. Bayes theorem is expressed below:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)}$$

Where: $P(D|E)$ = the probability of event D occurring, given that E is true.

$P(E|D)$ = the probability of event E occurring, given that D is true.

$P(D)$ = the probability of observing D.

$P(E)$ = the probability of observing E.

The Naïve Bayes classifier assumes that the attributes of the feature set are independent. Having made this simple assumption the likelihood of a feature set belonging to a certain class can be computed as the product of dimension-wise probabilities, which simplifies the calculation (Zaki and Meira, 2014). While the assumption that the variables in the feature set are independent might not technically be correct, the Naïve Bayes classifier is still able to produce accurate results and is most often used in the bag of words approach to text mining (see Section 3.5.6) (Han, Kamber and Pei, 2012). While the Naïve Bayes classifier assumes a specific feature probability distribution form for each class, the Logistic Regression classifier directly models the relationship probabilities. Thus, the assumption that each of the classifier makes about the underlying data is different.

In short, statistical classifiers map a set of feature variables to a class variable based on a probability distribution that is based on the training examples (Witten and Frank, 2005).

3.5.5.4 Neural Networks:

Neural networks are inspired by the human nervous system where nerve cells, called neurons, are connected to one another by synapses. Living organisms learn by strengthening the connections between neurons. The input data (from sense organs or other sources) trigger a neuron, which sends signals, via the synapses, to other neurons. The intensity of the signal corresponds to the strength of the synapse. This is mimicked in mathematical neural networks by using computations inside a node, which then triggers other nodes via its connections. The strength of the connections is weighted during the learning phase to mimic the strengthening of synapses (Aggarwal, 2015).

Neural networks are set up so that a node (n_i^k) corresponds to a variable in the feature set (x_i^k), for example one node for age and one node for salary. The nodes are arranged in layers where each input node is connected to each node in the following layer. The output layer can have a node for each class or a single output node. A node is assigned to a single class or a range of values is assigned to a class, for multi node and single node outputs respectively. The input will trigger the input nodes which transmits the input to the second layer via the weighted connections (w_i). The second layer will conduct a predefined arbitrary calculation, typically the sigmoid or logistic function. The results will be transmitted to the third layer via the weighted connections and so forth. The number of layers and number of nodes in a layer is specified by the investigators but a larger number of layers and nodes significantly increases the complexity of the network, and thus the calculation (Zaki and Meira, 2014; Aggarwal, 2015). A single layer neural network and a multi-layer neural network are shown below in Figure 3-6:

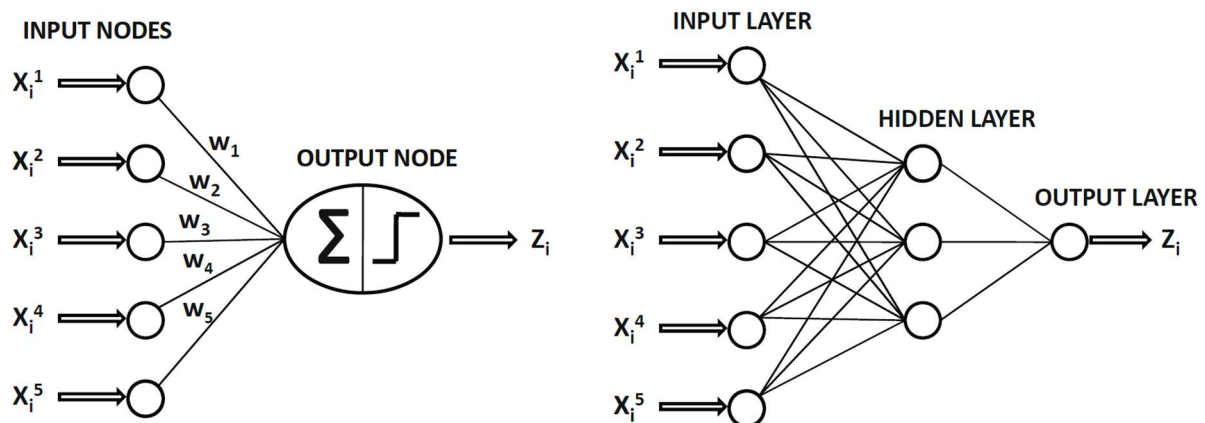


Figure 3-6: Neural Network Example (Aggarwal, 2015)

During the training phase the weighting for each connection is set to a random number. The input feature set is provided to the input nodes and compared against the class provided by the output node(s). For each iteration all the input data points are fed into the input nodes and

the output is recorded for each input set. The weighting between the nodes is adjusted slightly by a method called back propagated gradient decent. The weighting is adjusted to improve slightly for each iteration. The network therefore slowly learns the correct weighting for the training dataset (Zaki and Meira, 2014; Aggarwal, 2015).

3.5.5.5 K-Nearest-Neighbour:

First described in the 1950's, the K-Nearest-Neighbour approach to classification is based on comparing a given data point with training data points that are similar to it. The training data points have n number of features, meaning they can be represented and stored in an n -dimensional space. When an unknown data point is provided, the algorithm searches the n -dimensional space for the k number of training data points that are nearest to the unknown data point (Witten and Frank, 2005).

The 'closeness' of the data points are defined by a distance metric such as Euclidian Distance, given below for two data points with n features (Han, Kamber and Pei, 2012).

$$dist(\mathbf{X}_1, \mathbf{X}_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}$$

Features of varying scales, such as yearly income and age, must be normalised, as a difference in large features will overshadow the influence of a difference in small features. The K-Nearest-Neighbours classifier can accommodate categorical variables by examining if the data points are of the same category. If the categories are identical the distance is set as 0. If the categories are different the distance is set to maximum (1 if the data is normalised).

The unknown data is assigned the class that is most common amongst its k neighbours. The value of k is either assigned by the investigator or it is evolutionarily determined by testing a range of values (Han, Kamber and Pei, 2012; Zaki and Meira, 2014).

3.5.5.6 Support Vector Machines:

Support Vector Machines (SVM) were first presented by Vladimir Vapnik in 1992 and saw rapid adoption in the fields of data mining and machine learning due to their ability to accurately model linear as well as complex nonlinear decision boundaries (Witten and Frank, 2005; Han, Kamber and Pei, 2012).

Linear Support Vector Machines transform the original training data by linearly mapping it to a higher dimension. The SVM searches for the optimal linear separating hyperplane within this higher dimension. The hyperplane separates the two classes and is used as a decision boundary. The decision boundary is defined by a 'Support Vector' and its 'Margin'. The optimal hyperplane is found when the 'Margin' between the classes is maximised. Shown in Figure 3-7 is a two-feature linear SVM classification example with the optimal hyperplane i.e. the orientation of the 'Support Vector' allows for the 'Margin' to be maximised (Han, Kamber and Pei, 2012).

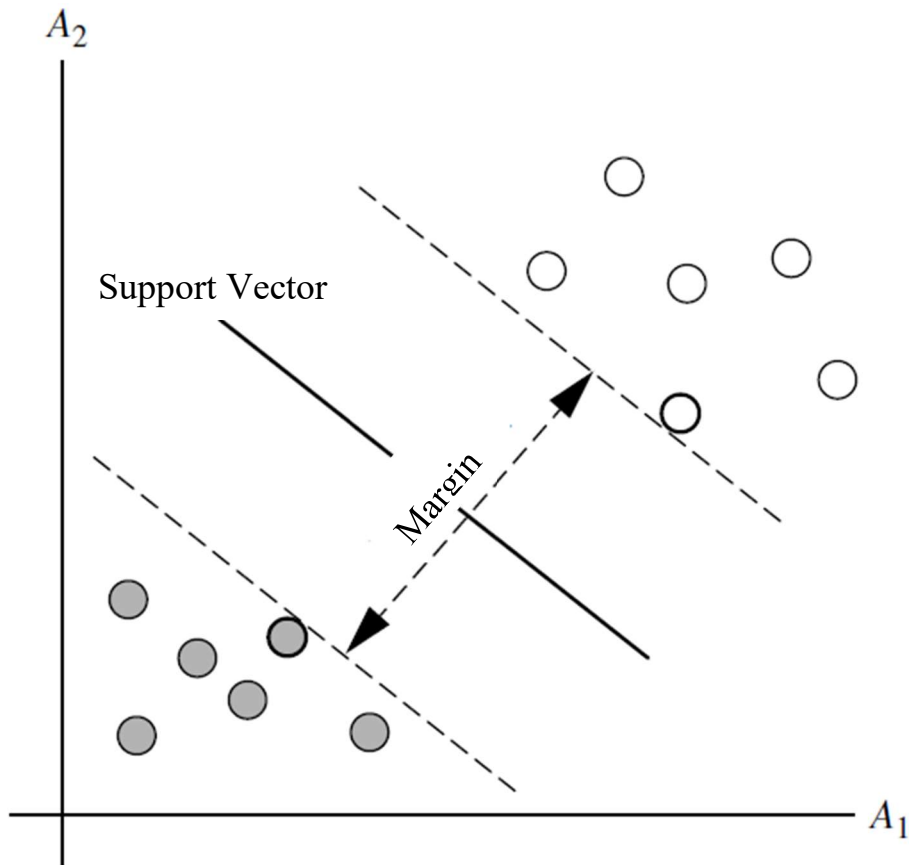


Figure 3-7: Support Vector Machine Example (Han, Kamber and Pei, 2012)

Linear Support Vector Machines are very accurate when the training data is linearly separable, as in Figure 3-6, but will fail to present a solution if the data is nonlinear. In Figure 3-8 a two-feature, linearly inseparable example is shown.

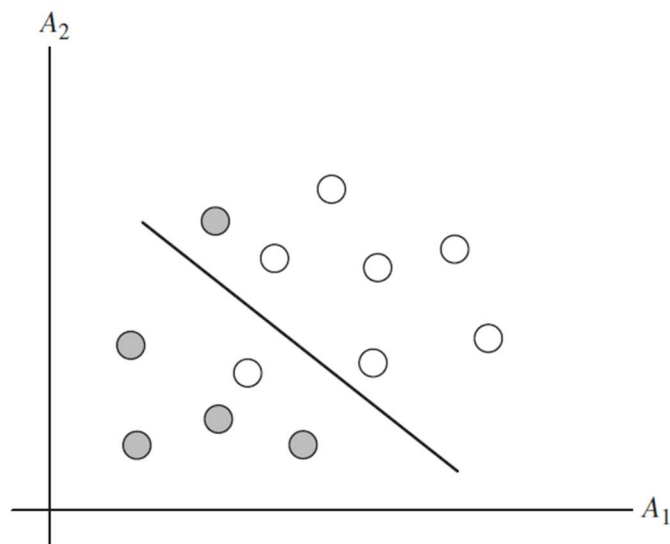


Figure 3-8: Linearly Inseparable Data SVM Example (Han, Kamber and Pei, 2012).

To enable Support Vector Machines to classify nonlinear data the projection of the data into higher dimensions is done using a nonlinear projection technique. The technique uses one of

a number of ‘kernel’ functions. These kernel functions usually need to be specified when implementing a Support Vector Machine (Han, Kamber and Pei, 2012; Zaki and Meira, 2014). Once the data has been projected into the nonlinear higher dimensions, the SVM searches for a linear decision hyperplane in the nonlinear space much like is shown in Figure 3-6.

3.5.5.7 Other Common Models:

Some other common classification model types are Stochastic Gradient Decent models, Instance-Based learning, Linear Discriminant analysis, semi-supervised learning, Ensemble Methods and others (Witten and Frank, 2005; Han, Kamber and Pei, 2012; Zaki and Meira, 2014; Aggarwal, 2015).

These methods all vary significantly in their implementation and approach to classification and regression modelling. The important common factor is that they all function on multidimensional data and can usually all be used on the same dataset, if the dataset is set up correctly.

3.5.6 Text Mining:

Text data is found in copious amounts in almost all domains. The data can be stored in a variety of forms such as documents, text on the internet, digital libraries and as text records of human speech. The set of unique words, or features, used in a piece of text is known as the lexicon and a collection of documents is known as a corpus. Text can either be treated as a multidimensional record or as a sequence but, due to the length of pieces of text and a lexicon that can often reach several hundreds of thousands of words, it is mostly viewed as a multidimensional record. The approach known as the ‘bag-of-words’ approach is used to represent text as multidimensional data (Zaki and Meira, 2014; Aggarwal, 2015).

The ‘bag-of-words’ approach uses a pre-processing phase where extremely common words, such as ‘and’, ‘to’ and ‘the’, are removed in a process known as stop word removal. Variations of the same words are consolidated in processes known as lemmatising and stemming. For example, ‘manage’, ‘managing’ and ‘managed’ will be reduced to their base form of ‘manage’ to reduce the number of unique words in the text. The processed text is then represented as an unordered set of unique words where the frequency of the individual words appearing in the text is associated with those words. Each text document is therefore represented as a vector where each row corresponds to a word with its frequency as the value in the vector. The size of the vector is set to the number of unique words that appear in all the text pieces that are being examined. These vector space representations of text are a type of multidimensional data which means that many of the classification methods discussed earlier can be used to classify the data (Aggarwal, 2015). Multidimensional text data has some unique features that must be considered and might require some modifications to the mining methods if they are to produce accurate results in a reasonable time frame.

A lexicon could be hundreds of thousands of words long, but a single document might only contain several hundred unique words. The vector for that document will therefore contain

mostly ‘zero’ entries. This is known as high-dimensional sparsity. The non-zero values will vary greatly between documents, which has a significant implication for distance calculations when comparing documents. When normalising data for non-text mining applications the data is typically given a value of between -1 and 1 but in text mining all the values are either zero or positive. The normalising techniques that are primarily used are as follows:

1. **Inverse document frequency:** words that appear frequently in text can be a source of noise in the data and can adversely affect similarity and distance computations. Inverse document frequency extends the concept of eliminating high frequency words in a softer way by assigning lower weights to words that appear more frequently.
2. **Frequency damping:** A damping function is applied to words that appear repeatedly in order to bring the frequencies between words that do appear in the text closer together. This is done because using the un-damped frequency for words that appear regularly has a significant biasing impact on the results of a similarity computation. Frequency damping can reduce the accuracy of clustering but has shown to increase the accuracy of classification results.

The fact that a word appears in a document is statistically more significant than a word not appearing in the document. Mining models must be cognisant of this fact and of the non-negativity of the data (Aggarwal, 2015). Models that work well in the high-dimensional, non-negative, sparse multidimensional data representations of text mining are support vector machines, Bayes classifiers and instance-based classifiers. Other classifiers experience computational efficiency issues that lead to long training and testing times.

3.5.7 Other Data Mining Types:

The data mining types mentioned in the mining and modelling section are the most common. Below are several other types of data mining algorithms that are generally created for mining a very specific data type, such as spatial data or graph data.

- **Mining data streams:** recent advances in data collection have allowed data to be continuously collected over time and at a rapid rate. These data streams introduce significant complications to the data mining process as the dataset is continuously being added to and increasing in size. These problems can be solved by the use of so called “Big Data” systems or other data sampling methods and algorithms suited to these situations (Han, Kamber and Pei, 2012).
- **Mining behaviour-context attribute data:** behaviour-context data is data that contains a behavioural attribute, a measurement, and contextual data, a time stamp or geographic location etc. Time-series data, spatial data, spatiotemporal data, and discrete data sequences are all examples of this type of data. Mining these types of data pose a unique challenge since the behavioural attribute cannot be considered in isolation. The contextual data has to be considered in conjunction with the behavioural data to produce meaningful results. Clustering algorithms, outlier

detection algorithms, and classification and regression algorithms therefore have to be modified significantly to account for this added data complexity (Aggarwal, 2015).

- **Mining graph data:** graph data is not data that is intrinsically behaviour-context data that can be represented on a graph, such as temporal data. Graph data is data that is represented on a graph and uses the edge of the graph to define the relationships between nodes. An example of this is a graphical representation of acetaminophen shown in Figure 3-9:

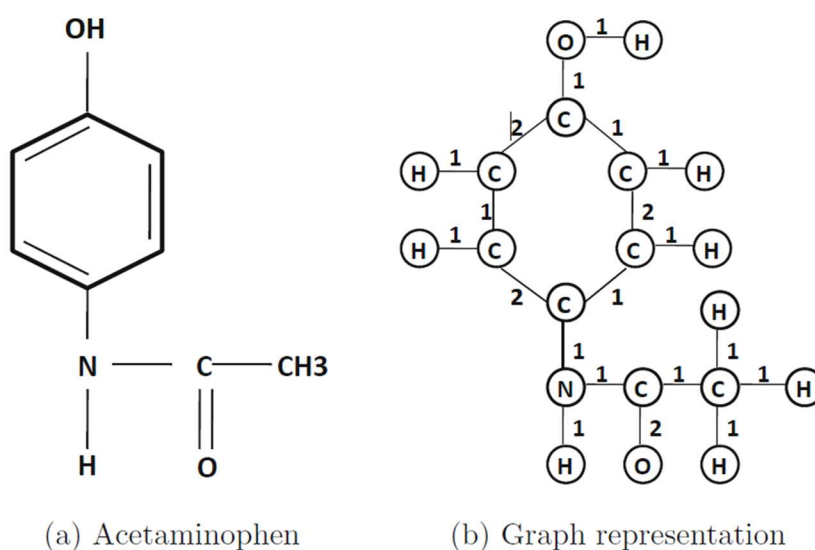


Figure 3-9: Acetaminophen and its graph representation (Aggarwal, 2015).

The mining of such data require new similarity and distance measures and therefore almost none of the methods mentioned in this chapter can be used for mining graph data (Aggarwal, 2015).

- **Mining web data:** web data comes in such a large variety of formats, quantity and speed that mining data from the web has been split into two broad categories: content-centric applications and usage-centric applications. These two applications use different overall approaches to mining their data, but even within each category there exists such a diverse number of techniques that it has become a field of research on its own (Han, Kamber and Pei, 2012; Zaki and Meira, 2014).

Many of the other types of data mining mentioned in this short section require significant modifications to the data mining algorithms mentioned in this chapter. These types of data mining tend to be beyond the reach of new data mining practitioners and require a more intricate knowledge of the field before an attempt can be made to implement such systems.

3.5.8 Mining and Modelling Applied to Construction Project Management Examples

The discussion of the data mining examples defined in Section 3.1.3 is continued in this Section. The goals of the applications were set in Section 3.2.2. The required data was collected in Section 3.3.4. The data was pre-processed in Section 3.4.5. In this section the mining and modelling of the applications is briefly discussed.

Mining and Modelling Procedure for Construction Cost Overrun Example: A cleaned and informative target dataset was produced from interviews that gathered performance and planning information about 77 construction projects. The authors selected a Support Vector Machine regression model (see Section 3.5.5.6 for the classification configuration) as their main model to conduct the analysis. They also selected a Multiple Linear Regression model, a *K*-Nearest-Neighbour regression model (see Section 3.5.5.5), a Decision Tree regression (see Section 3.5.5.1) model and a Neural Network regression model (see Section 3.5.5.4) to compare the efficacy of the SVM regression model.

Mining and Modelling for Prediction of Impact of an On-site Accident Example: A cleaned target dataset that contained features describing the on-site accident, the project monitoring information and the time overrun (%) would be prepared. A number of models should be selected for this application, such as a Support Vector Machine classifier, a Naïve Bayes classifier, a Decision Tree classifier and possibly a few more. This is done so that the most accurate model can be selected.

3.6 Validation and Evaluation:

Since the output from the data mining application in construction will be used as the basis for estimates, planning and decision-making, the financial ramifications mean that it is important that the results obtained from the data mining process are accurate. The process of determining the accuracy and reliability of the results is known as validation (unsupervised processes) and evaluation (supervised processes). The results of clustering and outlier analysis are therefore put through a validation process whereas the results from classification are put through an evaluation process. The main difference between the two is that unsupervised processes typically do not have a labelled testing dataset, while the supervised processes do (Farha Shazmeen *et al.*, 2013; Zaki and Meira, 2014). The validation process for clustering, outlier analysis, and classification and regression will be discussed in the three following sections. Association pattern mining has an inbuilt validation technique. The minconf and minsup variables that the investigator supplies to the algorithm, as discussed in Section 3.5.2, determine which rules are significant and need to be reported and are therefore not discussed here.

3.6.1 Cluster Validation:

The unsupervised nature of cluster analysis makes it difficult to validate clusters in real datasets as typically no external validation criteria are available (e.g. labelled test dataset).

This does not mean that it is impossible to validate the results of a clustering analysis. A number of methods and criteria have been defined to evaluate clustering analysis and can be split into internal, external, and relative validation criteria (Srivastava, 2014; Zaki and Meira, 2014).

1. **Internal:** These measure use criteria that are derived from the dataset itself. For example, a measure of the compactness of clusters and the separation between clusters can be obtained from the intra-cluster and inter-cluster distances.
2. **External:** These measures use criteria that are not inherent to the dataset. Synthetic testing sets, expert knowledge about the domain and the cluster that are expected are all external validation methods.
3. **Relative:** These measures use criteria that are derived from applying the same algorithm, with different algorithm parameters, to the same dataset and examining how the change in algorithm parameters affects the clustering outputs.

3.6.1.1 Internal Cluster Validation Criteria:

When clustering a real dataset, a labelled testing dataset does not typically exist. To overcome this issue and to validate the clustering obtained by an algorithm, internal criteria were defined that are derived from the dataset itself. The problem with internal validation criteria is that they are biased towards one algorithm or the other, depending on the base mechanism by which the clustering algorithm and the assessment criteria function (e.g. distance-based algorithms will be favoured by distance-based measures and density-based measures will favour density-based algorithms). Some commonly used internal validation criteria are as follows (Witten and Frank, 2005; Srivastava, 2014):

1. **Sum of square distances to centroids:** this measure is a distance-based criterion that sums the squares of the distances from the data points in a cluster to the centroid of that cluster. A smaller value for this measure is indicative of better clustering, but the measure will be biased towards distance-based algorithms.
2. **Intracluster to intercluster distance ratio:** This measure computes the ratio of the average distances of points in a cluster to the cluster centroid to the average distance between the centroids of the clusters. This measure is more sophisticated than the sum of square distances to centroids measure as it provides a normalised answer where lower values are better.
3. **Silhouette coefficient:** this is a measure of the average, minimum and maximum distances between points in a cluster and the average, minimum and maximum distances between clusters. The silhouette coefficient will range between -1 and 1 where 1 indicates clear clustering and -1 indicates 'mixing' of clusters.
4. **Probabilistic measures:** use a mixing model to determine the effectiveness of the clustering algorithm but require input as to the expected shape of the clustering.

The first three of these measures are distances-based measures and the fourth is a mixture model-based measure. The investigator must be cognisant of the fundamental method by which the chosen clustering model works so as to choose appropriate internal validation criteria.

The list above is not an exhaustive list of all the internal validation criteria but rather a brief explanation of some of the types of internal clustering criteria and why it is a difficult exercise. The exercise becomes even more difficult if the clusters are of an arbitrary shape and not circular or well separated shapes. All internal validation criteria make some underlying assumptions about the clustering within the dataset which might not necessarily be valid. This is a fundamental problem to all internal validation criteria and, currently, there are no satisfactory solutions to this problem (Han, Kamber and Pei, 2012).

3.6.1.2 External Cluster Validation Criteria:

True external validation criteria, such as a labelled test dataset rarely exist for a real clustering dataset. Methods exist to create a synthetic external dataset with known cluster labels. These methods must carefully mimic the conditions of the environment from which the real dataset is being drawn. They must be an accurate representation of what a real dataset might be like (e.g. number, shape and distribution of clusters should be realistic). This is critical as the algorithm will be tested on these synthetic external datasets to validate the algorithm and test its accuracy. Some external validation measures are as follows (Zaki and Meira, 2014):

1. **Purity:** This is a measure of the degree to which a cluster consists of only data points from the same class label. Note that one class could be split between multiple clusters and each cluster could have a high purity as they only contain one class label.
2. **Maximum matching:** This is a measure of how completely a single cluster encapsulates a single class label. This is a measure of how well one cluster catches all the data points from a single class.
3. **F-Measure:** Often there is a trade-off between purity and maximum matching. As a cluster becomes larger to catch more data points from a specific class (to increase its maximum matching) so does its chance of catching data points from other classes (reducing its purity). The F-measure is a measure of purity vs maximum matching.

Many other measures exist to measure the accuracy and reliability of a clustering algorithm when applied to an external labelled dataset but none of these measures will be reliable if the synthetic labelled dataset is not an accurate representation of the real datasets that are expected (Aggarwal, 2015).

3.6.1.3 Relative Validation Criteria:

When optimising a specific clustering algorithm, the input parameters (e.g. number of clusters, density of clusters etc.) are varied to determine which set of parameters is optimal

for a given set of data. Relative validation measures are employed to compare the output from varying the input parameters to the clustering algorithm to find the optimal input parameters. Relative validation criteria are only used when a labelled test dataset does not exist. Many of the internal validation criteria can be modified to serve as relative validation criteria. Relative validation criteria aim to determine the stability of clusters and the clustering tendency (Witten and Frank, 2005).

Cluster stability refers to the sensitivity of clustering results obtained when varying input parameters. The main idea is that clusters obtained from an algorithm applied to different subsets of the main dataset should be similar or stable. A single dataset will therefore be broken up into several subsets of data. The algorithm will be applied to the subsets of data and if similar clusters are obtained from the analysis of the different subsets, then the algorithm produces stable clusters. But if the clusters vary largely from one subset to the other, then the algorithm produces unstable clusters (Hawkins *et al.*, 2002; Zaki and Meira, 2014).

Clustering tendency or ‘clusterability’ refers to whether the dataset can actually be grouped in any meaningful way or if the data does not have any internal structure. This is a tricky task as the definition of a cluster differs from one clustering algorithm to the next. Even if the definition is fixed, the task will still be difficult but nevertheless useful. Several methods exist to determine the clustering tendency of a dataset, however these methods do not tell the investigator how many clusters there might be, only that the dataset contains some meaningful clusters (Zaki and Meira, 2014).

3.6.2 Outlier Validation:

Cluster analysis and outlier detection are complementary processes, but the validation of outliers cannot be defined in a similarly complementary way. Clustering and Outlier analysis are both unsupervised learning problems, therefore many of the problems of validating clusters also exist for validating outliers but to a greater extent. As for cluster analysis, external labelled datasets with which to evaluate an unsupervised learning process rarely exist, but synthetic datasets can be generated. Internal validation methods were defined for cluster analysis that used measures derived from the dataset as a method to evaluate the clustering, but internal validation methods are rarely used for outlier validation (Aggarwal, 2015).

3.6.2.1 Internal Outlier Validation:

Internal validation methods are known to be flawed when used in cluster analysis. The bias of the internal validation method criteria towards an algorithm of the same base type is significantly exaggerated if applied in outlier validation. This is because the internal clustering methods are good at identifying the body of the clusters and will tend to agree with each other even if they use different underlying methods. The exact clustering method used will affect the interpretation of the fringe cases of the dataset. These fringe cases are points that lie in an ambiguous zone where it is difficult to assign either a cluster label or an outlier

label. A single data point might be assigned the label of an outlier by one algorithm whereas a different algorithm might deem it to be normal. If the internal validation criteria utilises the same base function as the first algorithm it is likely to agree with the label. Therefore, the internal validation criteria will be extremely biased towards an outlier analysis method that uses a similar underlying method. This makes internal validation methods for outlier validation basically untenable (Hawkins *et al.*, 2002; Aggarwal, 2015).

3.6.2.2 External Outlier Validation:

External validation methods are the measures that are typically used when validating outlier analysis. Real labelled datasets rarely exist and therefore synthetic external datasets are typically generated to evaluate outlier analysis. Typically, outlier analysis uses a numerical value to determine a data point's 'outlier-ness'. If the threshold for labelling a point as an outlier is set too conservatively, then the algorithm can miss true outliers (False-Negative), but if it is set too low then the algorithm can incorrectly label a point as an outlier (False-Positive). Therefore, a trade-off exists between false-positives and false-negatives. This threshold cannot be known exactly for a real dataset, but the trade-off between false-positives and false-negatives can be graphed. The graph can be used to compare various algorithms. The Receiver Operating Characteristic (ROC) curve is one such graph that graphs the false-positive rate against the true-positive rate for various cut-off values when the algorithms are run on a labelled synthetic dataset. The 'lift' above the diagonal line is a qualitative measure of the accuracy of an algorithm. Otherwise, the area under the ROC curve can be used as quantitative measure of the algorithm accuracy (Hawkins *et al.*, 2002). An example of a ROC curve is shown below in Figure 3-10 where it can be seen that algorithm A achieves a better 'lift' than algorithm B and is therefore a better algorithm for outlier analysis on this particular dataset.

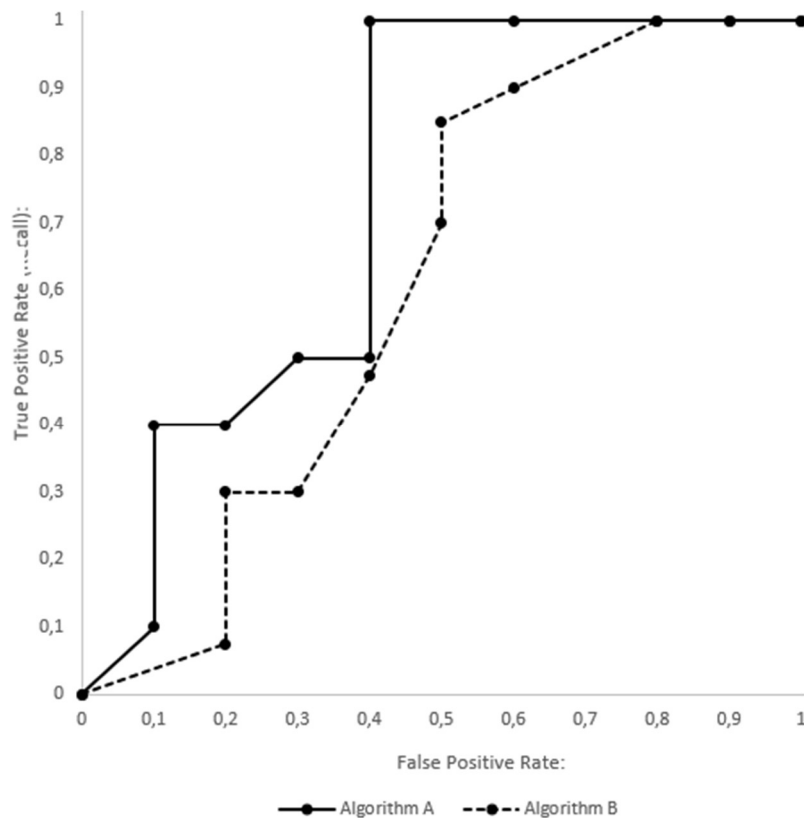


Figure 3-10: ROC curve for Outlier Analysis Validation

3.6.3 Classifier and Regression Evaluation:

The evaluation step in the data mining process is critical if one wishes to have accurate and dependable results. Determining the classifier's effectiveness, comparing different classifiers to choose the best model for a specific set of data, and parameter tuning are all part of evaluating a classifier (Liaw and Wiener, 2002; Srivastava, 2014). When evaluating a classifier, a labelled set of test examples is required that acts as the 'ground truth' against which the classifier is tested. The evaluation process for a classifier, while simpler and more straightforward than the evaluation process for clustering and outlier analysis, can still produce results that overestimate or underestimate the validity of a classifier. The performance measures for classifiers and regression models are discussed below along with the classification and regression evaluation methodology.

3.6.3.1 Classifier Performance Measures:

There are several common measures that are used to evaluate the accuracy of a classifier. These measures combined with the classifier evaluation methods (Section 3.6.3.3) discussed later allows the investigator to accurately compare different classification algorithms. To aid the explanation of the classifier performance measures an example is used. Table 3-1 shows the predictions made by a classifier when tested on a dataset containing 10 data points for each of the three classes in the dataset based on an example from Zaki and Meira (2014).

Table 3-1: Classification assessment example

	True:			
Predicted:	Class 1:	Class 2:	Class 3:	Sum:
Class 1:	$n_{11} = 10$	$n_{21} = 0$	$n_{31} = 0$	$m_1 = 10$
Class 2:	$n_{12} = 0$	$n_{22} = 7$	$n_{32} = 5$	$m_2 = 12$
Class 3:	$n_{13} = 0$	$n_{23} = 3$	$n_{33} = 5$	$m_3 = 8$
Sum:	$n_1 = 10$	$n_2 = 10$	$n_3 = 10$	$n = 30$

Of the thirty data points (n) labelled by the classifier, twelve values were labelled as Class 2 (m_2). Seven of the twelve data points labelled by the classifier as Class 2 were correctly labelled (n_{22}) and five of the twelve incorrectly labelled as Class 2 as they were actually Class 3 (n_{32}). Below, the classifier performance measures are explained and calculated for the table above (Witten and Frank, 2005; Srivastava, 2014).

- **Error Rate:** is a measure of the misclassification probability of the classifier for all the classes in the dataset. Therefore, if the error rate is lower, the classifier is better. The error rate is calculated as the fraction of predictions that were incorrect on the total dataset.

$$\text{Error Rate} = \left(\frac{n_{12} + n_{13} + n_{21} + n_{23} + n_{31} + n_{32}}{n} \right)$$

$$\text{Error Rate} = \left(\frac{0 + 0 + 0 + 3 + 0 + 5}{30} \right) = 0.267$$

- **Accuracy:** is a measure of the correct prediction probability of the classifier for all the classes in the dataset. Therefore, if the accuracy is higher, the classifier is better. The accuracy is calculated as the fraction of prediction that were correct on the total dataset.

$$\text{Accuracy} = \left(\frac{n_{11} + n_{22} + n_{33}}{n} \right)$$

$$\text{Accuracy} = \left(\frac{10 + 7 + 5}{30} \right) = 0.733$$

- **Precision:** is a measure of the class specific accuracy or “exactness” of the classifier. Therefore, if a class precision is higher, the classifier is better at classifying that class. The precision for a class is calculated as the correct predictions for that class over all the values predicted to be in the class.

$$\mathbf{Precision} = \left(\frac{\mathbf{n_{22}}}{\mathbf{m_2}} \right)$$

$$\mathbf{Precision} = \left(\frac{\mathbf{7}}{\mathbf{12}} \right) = \mathbf{0.583}$$

- **Recall/Coverage:** is a measure of a classifier’s ability to identify all the points in a class. Therefore, the higher the recall for a class, the better the classifier. The recall for a class is calculated as the number of correct predictions for a class over all the values that are in that class.

$$\mathbf{Recall} = \left(\frac{\mathbf{n_{22}}}{\mathbf{n_2}} \right)$$

$$\mathbf{Recall} = \left(\frac{\mathbf{7}}{\mathbf{10}} \right) = \mathbf{0.7}$$

- **F-Measure:** There is often a trade-off between the recall of a classifier and the precision of a classifier. For example, by labelling all testing points to be in one class, the classifier’s recall for that class will be equal to 1. However, the precision of the classifier for that class will be very low. This can be reversed where the classifier only classifies a few points as one class, where the classifier has the most confidence. The classifier will get an extremely high precision value for that class but a very poor recall value. This trade-off between precision and recall is quantified in the F-measure and the aim is to achieve as high an F-measure as possible. The F-measure is quantified as follows:

$$\mathbf{F} = \frac{\mathbf{2 \times Precision \times Recall}}{\mathbf{Precision + Recall}}$$

$$\mathbf{F} = \frac{\mathbf{2 \times 0.583 \times 0.7}}{\mathbf{0.583 + 0.7}} = \mathbf{0.636}$$

- **ROC Curve:** The receiver operating characteristic (ROC) curve is an evaluation method that is used when the output from the classifier is a numerical score or when there are two classes in the dataset. The ROC curve indicates how much the data point fits in a specific class. A numerical threshold (t) for labelling the data point as a specific class is typically specified as an input criterion for a model. The ROC curve is defined by plotting the false positive rate on the x-axis and the true positive rate on the y-axis for varying threshold values (t = 0 to t = 100). A completely random classification method is expected to produce a diagonal line. The *lift* above the diagonal line provides the user with an idea of the accuracy of the classifier. A more concrete quantitative value is the area under the ROC curve, where a greater value

means the classifier is more accurate. The area under a ROC curve is a useful method for directly comparing classifiers as the classifier with the highest score is the most appropriate for the application. An example of an ROC curve is shown in Figure 3-9 in Section 3.6.2.2.

3.6.3.2 Regression Performance Measures:

The options of assessing the effectiveness of regression models are more limited than the classifier performance measures. The Coefficient of Determination, or the R^2 -statistic, is the performance measure most commonly used for linear regression models. The Coefficient of Determination (R^2), the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are calculated as follows (Lee *et al.*, 2011; Han, Kamber and Pei, 2012; Srivastava, 2014):

$$R^2 = 1 - \frac{\sum_{i=1}^n (x_i - y_i)^2}{\sum_{i=1}^n (x_i - \bar{x}_i)^2};$$

$$MAE = \frac{1}{n} \sum_i |x_i - y_i|;$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2};$$

Where: n = the number of data points.

x_i = actual data value.

\bar{x}_i = mean of actual data value

y_i = predicted value

The R^2 statistic ranges between 0 and 1 for linear models, with 1 being a perfect model. If the data has a high dimensionality, the adjusted R^2 statistic is a more accurate measures, where d is the number of features and n is the sample size (Witten and Frank, 2005; Lee *et al.*, 2011; Srivastava, 2014).

$$R^2 = 1 - \frac{(n - d - 1)SS_{res}}{(n - 1)SS_{tot}}$$

In the case of a nonlinear model the R^2 statistic can be very misleading, or even negative. For nonlinear models the MAE and RMSE are used as measures of the error with a perfect model achieving a score of zero for both these measures (Lee *et al.*, 2011; Aggarwal, 2015).

3.6.3.3 Classifier and Regression Evaluation:

The evaluation methods discussed in this section use a testing dataset and the above-mentioned classifier evaluation criteria to determine how good a classifier is. The methods discussed below aim to remove the possibility of over estimating or underestimating the accuracy of a classifier. They are designed to provide a dependable quantification of the accuracy of a classifier.

- **Holdout:** The holdout method is the most common approach to classifier evaluation. The holdout approach randomly splits the labelled dataset into two subsets. Two thirds or even three quarters of the total labelled dataset is randomly assigned to the training dataset whereas the remaining data is assigned to the testing dataset. The classifier is trained on the training set and evaluated, with the evaluation measure discussed above, on the testing set. This splitting, training and testing process can be repeated multiple times to determine the variance of the error estimates. A problem with the holdout evaluation method is that underrepresented or overrepresented classes in the training or testing dataset will bias the evaluation criteria. For example, if a dataset of 1000 points contains 950 points from one class and 50 from another class and the data is randomly split 70-30 into the training and testing dataset, then the testing dataset could conceivably end up with few or no data points of the sparse class. A reasonable solution to this problem is to ensure that the training and testing dataset are representative of the data environment, where the classes are present in the two dataset in the same ratio that they are present in the overall labelled dataset (Witten and Frank, 2005; Aggarwal, 2015).
- **K-Fold Cross-Validation:** Cross-validation is an evaluation method that divides the labelled dataset into K equal-sized parts, known as *folds*. A typical number of folds is $K = 10$. One fold is reserved as a testing dataset and the classifier is trained on the remaining folds. The classifier is evaluated with the evaluation criteria discussed above and then retrained where a different fold is reserved as the testing set and the rest are again used as the training set. This process is repeated until each fold has been used as a testing set. The evaluation criteria are averaged over the number of folds to determine the overall accuracies. Cross-validation tends to provide a highly representative estimate of the model accuracy. This entire process can be repeated a number of times in order to obtain the mean and variance of the evaluation criteria (Witten and Frank, 2005; Zaki and Meira, 2014).
- **Bootstrap Resampling:** Bootstrap resampling is an evaluation method that is typically only used when the labelled dataset is small. The method creates a training dataset that is the same size as the original dataset by randomly selecting data points, with replacement, from the labelled dataset. The result is a training dataset that is of the same size as the original labelled dataset, but with certain data points being duplicated and certain data points being left out. The model is trained on the constructed training dataset and the model is evaluated using the original labelled dataset. This method is a highly optimistic method and can produce positively skewed results of the classifier's accuracy because of the large overlap of training and testing data (Zaki and Meira, 2014; Aggarwal, 2015).

Once the evaluation process has been completed and the performance measures obtained, the investigator must determine if the accuracies obtained are acceptable, or if further investigation must be done to increase the accuracies. The range of acceptable values will differ largely between applications and sectors. For example, the acceptable accuracy for a machine learning model designed to examine biopsies and determine if they are benign or cancerous will require very high accuracies (98-99.99% accuracy rate). Whereas sentiment analysis of movie review will typically be acceptable if the accuracy is higher than 60%. The investigator needs to make an independent decision about whether the performance measures obtained are acceptable (Han, Kamber and Pei, 2012). If the investigator decides the measures are deemed unacceptable, the investigator may then go back through the data mining process adding data, changing pre-processing techniques or altering the data mining model.

The above processes can also be used for evaluating the accuracy of a regression model. A common mistake is that the investigator uses the testing dataset to make choices about which algorithms to use and to adjust the input parameters of the algorithm. This is dangerous as knowledge and considerations about the testing dataset is being implicitly used by the investigator to have the algorithm perform better on the testing set. This can lead to overestimating the accuracy of the algorithm (Witten and Frank, 2005; Aggarwal, 2015).

3.6.4 Validation of Construction Project Examples

The discussion of the data mining examples defined in Section 3.1.3 is continued in this section. The goals for the applications were set in Section 3.2.2. The required data for the applications was collected in Section 3.3.4. The data was pre-processed in Section 3.4.5. The mining and modelling was conducted in Section 3.5.8. In this section the validation and evaluation of the results of the application is conducted and the examples concluded.

Validation Procedure Applied to Cost Overrun Example: 5 regression algorithms were selected to model to the problem of predicting cost overrun on construction projects in South Korea. The performance measures used by the authors were the correlation coefficient (R) (see Section 3.6.3.3), the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) (see Section 3.6.3.2).

For high prediction accuracy the correlation coefficient (R) should be close to 1 and the MSE and RMSE should be close to 0 (Lee *et al.*, 2011). The authors utilised these performance measures within a ten-fold cross validation process. The results for the five regression models over the ten folds are shown in Table 3-2.

Table 3-2: Results of Cost Overrun Prediction Example (Lee et al., 2011):

Model:	R	MAE	RMSE
Support Vector Machine	0.8	4.72	6.61
Multiple Linear Regressions	0.62	6.93	8.60
K-Nearest-Neighbour	0.78	5.31	6.86
Decision Tree	0.51	7.27	9.38
Neural Network	0.61	6.83	8.59

The Support Vector Machine was best able to predict the cost overrun of a construction project. The high correlation coefficient of 0.8 and the relatively low MAE and RMSE of 4.72 and 6.61 show that the model is able to reliably predict the cost overrun of a construction project. The model can be put into use by having contractors or project managers answer the questionnaire used in the data acquisition phase and provide the results to the Support Vector Machine model to predict the potential cost overrun on their project. This information can then be used to assign resources to potentially troubled projects.

Validation of Results for Prediction of Impact of an Accident on site: A number of algorithms may be selected for modelling the prediction of the time overrun caused by an accident or safety issue on-site. The performance measures used for the investigation should be the Accuracy and Recall (see Section 3.6.3.1). To determine the overall and the class specific accuracy of the classifier these performance measures should be used during a ten-fold cross validation process. The accuracy obtained must be interrogated to determine if it is adequate. If the accuracy of the best performing model is adequate, then the model can be used by construction project in the risk identification and analysis phase.

3.7 Conclusion:

Chapter 3 discussed the data mining process defined in Chapter 2 by drawing on information from the construction project environment and three leading data mining textbooks. Chapter 3 provided a clear description of how a data mining application can be developed and implemented. The discussion of each step introduces a number of possibilities a data mining practitioner might utilise to complete each step. For example, the Data Acquisition step discussed a number of possible data sources, and the Mining and Modelling step discussed

several data mining types that could be used by an investigator. The exact methods used to implement the data mining are not discussed as these will differ from application to application, but sufficient information is provided to inform construction industry professionals of the many options available. A summary of Chapter 3 is provided in Appendix 2.

4 Data Mining Resources:

4.1 Introduction:

Chapter 4 focuses on the available data mining resources that enable the application of data mining without requiring the investigator to implement the data mining algorithms discussed in Chapter 3 from first principles. These resources are then used in the data mining application presented in Chapter 5. The discussion of these resources fulfils the third of the four objectives for the study. The base of knowledge provided in the previous chapter, combined with the implementation resources presented in this chapter will enable trained construction sector personnel, without a data mining background, to implement a data mining application.

The toolkits and software packages presented in this chapter are selected due to their comprehensive nature and their user-friendliness. The research defined a comprehensive data mining package as containing the following:

1. Methods for the pre-processing of collected data into a clean, normalised and informative target dataset.
2. Methods to perform classification and regression modelling along with the required performance criteria and evaluation processes.
3. Methods to perform unsupervised learning such as clustering and outlier analysis along with the required validation criteria.

The availability of text mining capabilities will be considered a bonus but will not be required.

Since data mining and machine learning are fields of statistics and computer science, these resources tend to assume some knowledge of programming, usually in the Python language or the R statistical language. There are resources that allow an investigator to create data mining applications without the practitioner writing any code, but these tend to be paid for products, whereas the Python and R packages are free to use under the open-source BSD licence agreement (Loper and Bird, 2004; Pedregosa *et al.*, 2011).

Two data mining toolkits are available for investigators with experience of programming in Python:

- **Scikit-Learn:** a simple, efficient, user-friendly and powerful toolkit for data mining and data analysis built on common Python modules.
- **Natural Language Toolkit (NLTK):** an easy-to-use platform for symbolic and statistical processing of natural human language.

A toolkit for data mining in the R statistical language is available:

- **Machine learning in R (mlr):** provides the infrastructure for the application of a variety of data mining algorithms in a user-friendly manner.

A free-to-use software package is available:

- **Orange:** contains a collection of data visualisation tools and algorithms for data mining and analysis with a visual programming interface for ease-of-use.

A paid-for software package is available:

- **RapidMiner:** is a data science and data mining platform developed to provide an integrated visual environment for data preparation, machine learning, and predictive analytics.

These resources are evaluated by comparing the functionality they provide to the requirements of the data mining examples discussed in Chapter 3 (see Section 3.1.3). Recommendations are made regarding which resource might be the most suitable depending on the requirements of the application and the expertise of the investigator.

The list of resources is not exhaustive as there are many other comprehensive data mining toolkits. A good understanding of the data mining process, provided by Chapter 3, and a willingness to learn new skills will enable a novice in the data mining field to gather meaningful insights from, and produce valuable models of, their data.

4.2 Scikit-learn:

Scikit-learn is a Python machine learning module that was born out of a Google ‘Summer of Code’ project. The library has been active since 2007 and has become one of the most used data mining and machine learning libraries. Scikit-learn specialises in providing a wide variety of methods in a user-friendly way (Pedregosa *et al.*, 2011). This user-friendly library allows the novice investigator to set up a target dataset and then to select multiple algorithms to use as and test for appropriate application without changing the target dataset (Pedregosa *et al.*, 2011). Some of the libraries within the Scikit-learn module are:

- **Pre-processing:** Provides a large number of algorithms that enable the investigator to standardise and normalise their dataset, scale data that contains outliers, encode categorical variables, transform continuous values into discrete bins and more. The functionality allows the investigator to complete the pre-processing phase to produce a good target dataset.
- **Dimensionality Reduction:** Provides tools and methods to determine which features in a dataset are the most informative and should be included in the target dataset. Dimensionality reduction tools such as single value decomposition are provided. This library can be used during the pre-processing step in the data mining process.
- **Clustering:** Provides 9 different algorithms types, each with several sub-categories, to perform automatic clustering on a dataset. A valuable overview table for each algorithm is provided to assist the investigator in selecting appropriate algorithms and

a clustering chart is also provided to compare how the algorithms classify 5 mock datasets (see Figure 4-1). The library can perform basic outlier analysis.

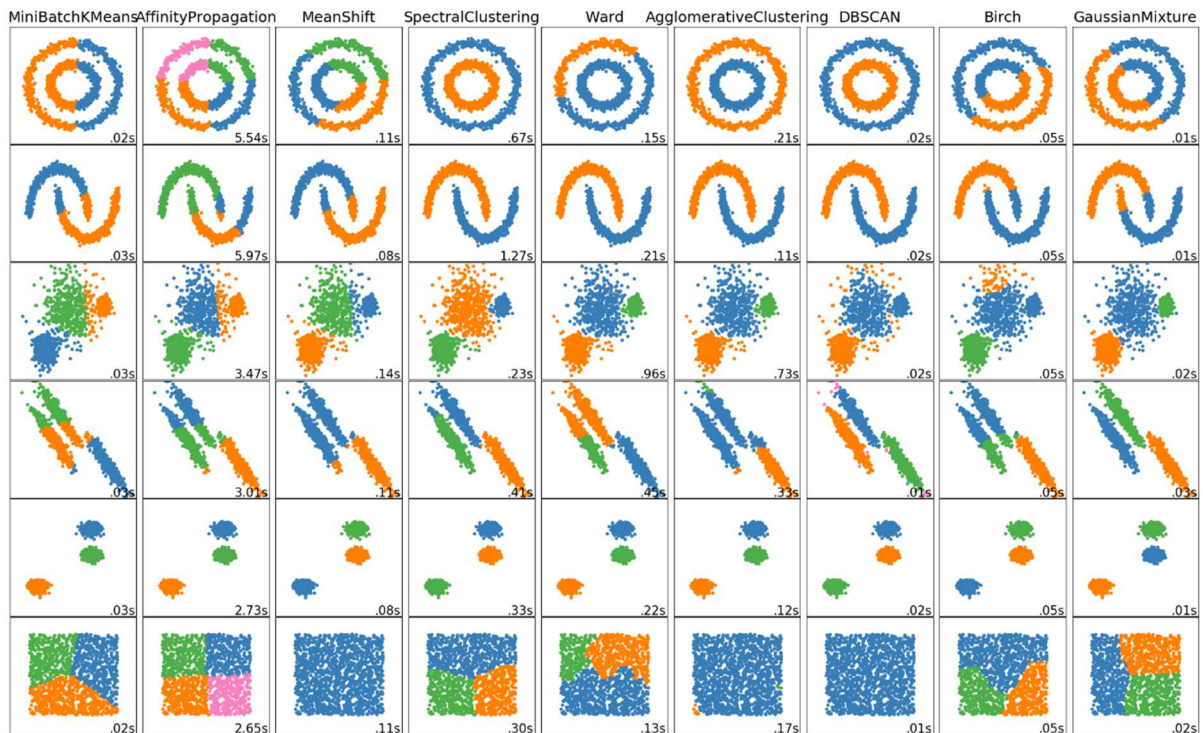


Figure 4-1: Visual representation of 9 clustering algorithms provided by Scikit-learn applied to 6 different unlabelled datasets (Pedregosa et al., 2011)

- **Classification:** The classification package contains a large number of different algorithms that can be used for data classification. The input and output of the algorithms are the same format allowing greater ease in comparing algorithms.
- **Regression:** Nearly all the algorithms that are used in the classification package have been adapted for use in the regression package.
- **Model Selection:** Scikit-learn provides a number of tools to perform algorithm validation and evaluation. These tools include automatic k-fold cross-validation and parameter tuning. The parameter tuning tools allow the investigator to fine-tune the input parameters to achieve the maximum accuracy possible. Tools are also provided that allow algorithm comparison so that the best algorithm can be selected in a diligent and thorough manner.

If the investigator is a complete novice and has no idea where to start their investigation, Scikit-learn provides an algorithm cheat sheet that will guide the investigator to an algorithm they should be able to implement. The cheat sheet is shown in Figure 4.2:

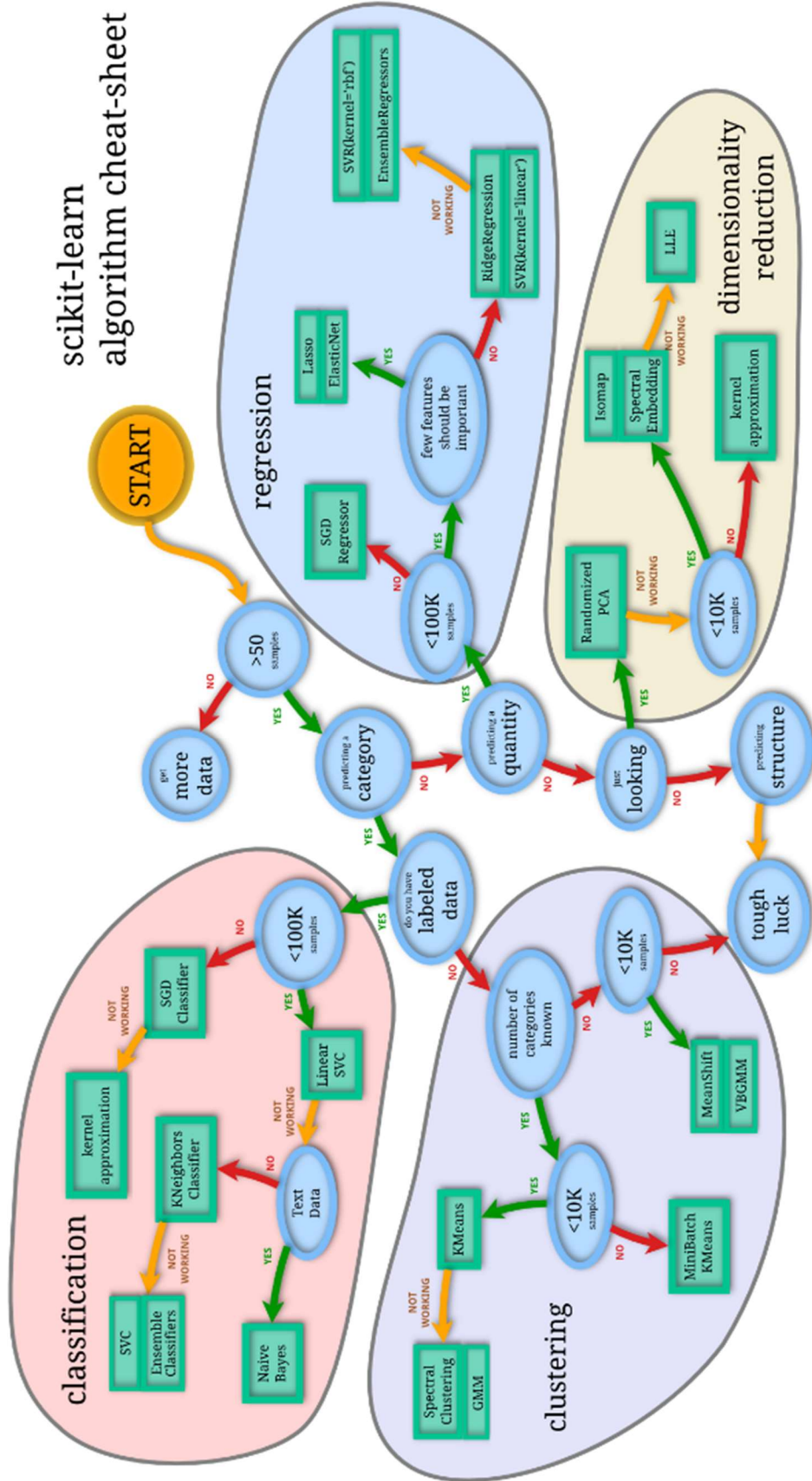


Figure 4-2: Scikit-learn algorithm cheat sheet (Pedregosa et al., 2011)

Scikit-learn can be found at: <http://scikit-learn.org/stable/>

4.3 Natural Language Toolkit:

The Natural Language Toolkit (NLTK) is a Python package that allows users to process human language data. NLTK was developed in the Department of Computer and Information Science at the University of Pennsylvania and was initially released in 2001. The toolkit is intended to support research in natural language processing and has become a teaching and research tool at 32 universities. A textbook was published to accompany the toolkit and thoroughly explains the different libraries, methods, and their implementation (Loper and Bird, 2004)

Text mining (see Section 3.5.6) uses many unique data pre-processing techniques and requires modified data mining algorithms. NLTK provides many of these techniques and modified algorithms in an easy-to-use format. The toolkit provides libraries with the following capabilities:

- **Tokenisation:** Splitting a text string into individual entities (words, punctuation, numbers etc)
- **Stemming:** Reducing inflected words to their base form.
- **Lemmatising:** Grouping inflected words and representing them in their base form. Lemmatising is a more sophisticated approach than stemming but the results are similar.
- **Tagging:** Each token is assigned a label using part-of-speech tagging.
- **Information Extraction:** Perform information extraction using named entity recognition and other algorithms.
- **Parsing:** Use tagging to construct a sentence parse tree, as shown in Figure 4-3 for the sentence, “The quick brown fox jumped over the lazy dog”



Figure 4-3: Parsing sentence structure using NLTK (Loper and Bird, 2004).

- **Wrapping:** NLTK provides several ‘wrappers’ that allow processed text data to be used by other machine learning toolkits, such as Scikit-learn.
- **Classification and Clustering:** Provides functionality to perform document classification and clustering using a number of underlying algorithms.

The toolkit also provides over 50 corpora of books, lexical resources and other text files for use when testing text processing applications. The accompanying textbook provides a hands-on tutorial for all the methods contained in the module. An active discussion forum exists where users engage each other and provides support for users with a basic knowledge of Python programming to use the module. The toolkit has user-friendly appeal for novice users but also contains advanced methods that make it suitable for experienced programmers (Bird, Klein and Loper, 2009). The ‘wrapping’ function that allows cross-compatibility with other machine learning modules enables a user to apply sophisticated data mining algorithms to text data

4.4 Machine Learning with R (mlr):

The machine learning with R (mlr) package for the R statistical programming language provides a generic, object-oriented and extendable framework for a number of machine learning applications. The package is aimed at practitioners who wish to quickly apply data mining and machine learning as well as researchers who wish to implement, benchmark, and compare new and prewritten methods in a controlled environment (Bischl *et al.*, 2016).

The mlr package provides the following capabilities in a user-friendly environment (Bischl *et al.*, 2016):

- **Tasks and Learners:** The data and other relevant information is encapsulated within tasks. These tasks support a number of machine learning ‘learners’. These include regular and multiclass classification, regression, clustering and survival analysis. There are currently 82 classification, 61 regression, 13 survival and 9 clustering learners available.
- **Evaluation and Resampling:** The framework provides the capability to validate the output from the ‘learners’. Currently 46 performance measures are usable with subsampling, bootstrapping, and cross-validation. The framework also supports clustering validation criteria.
- **Tuning:** mlr provides the ability to automatically optimise the parameters for all the ‘learners’ and several other pre-processing functions.
- **Feature Selection:** mlr supports wrapper and filter approaches to feature selection and allows for the visualisation of the improvements in the features selected.
- **Wrapper Extensions:** The mlr framework provides wrappers that allow the user to implement custom scripts throughout the data mining process.
- **Benchmarking and Parallelisation:** The computational efficiency of ‘learners’ can be evaluated to rank them from fastest to slowest.

- **Properties and Parameters:** The mlr tasks have the ability to respond to queries by the user on the type of data it supports, its implementation, the required parameters and more.

There are a number of other machine learning packages for R, but mlr provides a comprehensive and user-friendly package that enables the user to implement both beginner and advanced data mining applications (Bischl *et al.*, 2016).

4.5 Orange:

Orange is an open-source data mining, data analysis and data visualisation toolkit with a visual programming interface. Orange started development at the University of Ljubljana in Slovenia. The software package continued development and currently has an international team of developers. Its intuitive visual programming interface has led to Orange being used at Universities and data mining training courses across the world (Demšar *et al.*, 2013). The machine learning, data mining and data visualisation tools the toolkit provides are:

- **Data:** The toolkit provides a number of data collection, extraction and handling techniques. The toolkit supports data from a wide variety of sources. The ability to identify and remove outliers from the data, to impute missing values, to discretise continuous data into bins and to perform all other necessary pre-processing steps is provided within this module.
- **Visualise:** The toolkit provides the capability to visualise the data at any stage of processing in a wide variety of graphs and visual formats.
- **Model:** The toolkit provides a number classification and regression algorithms. These algorithms are all accompanied by specialised visualisation techniques.
- **Evaluate:** The ability to test and evaluate the results of the data modelling is provided.
- **Unsupervised:** A number of clustering algorithms and clustering validation procedures are provided.
- **Specialised Add-ons:** Specialised add-ons have been provided that extend the capability of the toolkit but might not be applicable to every user.
 - **Associate:** This add-on provides the ability to perform association pattern mining and frequent itemset mining.
 - **Bioinformatics:** This add-on focuses on data mining of genes and other bio-information.
 - **Data Fusion:** This add-on allows for the reduction and joining of datasets, such as with single value decomposition.
 - **Image Analytics:** This add-on extends the ability of the toolkit to process and model a number of image related applications.

- **Education:** The developers of the toolkit supplied this add on as a hands-on tutorial and to enable teaching of data mining at Universities.
- **Text Mining:** The ability to pre-process text and to perform classification on text are provided in this add-on. The add-on also provides a number of large text corpora.
- **Time Series:** Time series mining requires many unique pre-processing steps and data mining algorithms. This add-on provides the required techniques to allow the user to perform time series mining.

The Orange toolkit provides detailed explanations and implementation guides of the 100+ methods the toolkit contains. The visual programming interface reduces the barrier to entry for data mining by removing the need for the user to be able to program, either in Python or in R. Figure 4-3 below shows the visual programming used on the cross-validation and scoring of three classifiers (classification tree, SVM, AdaBoost).

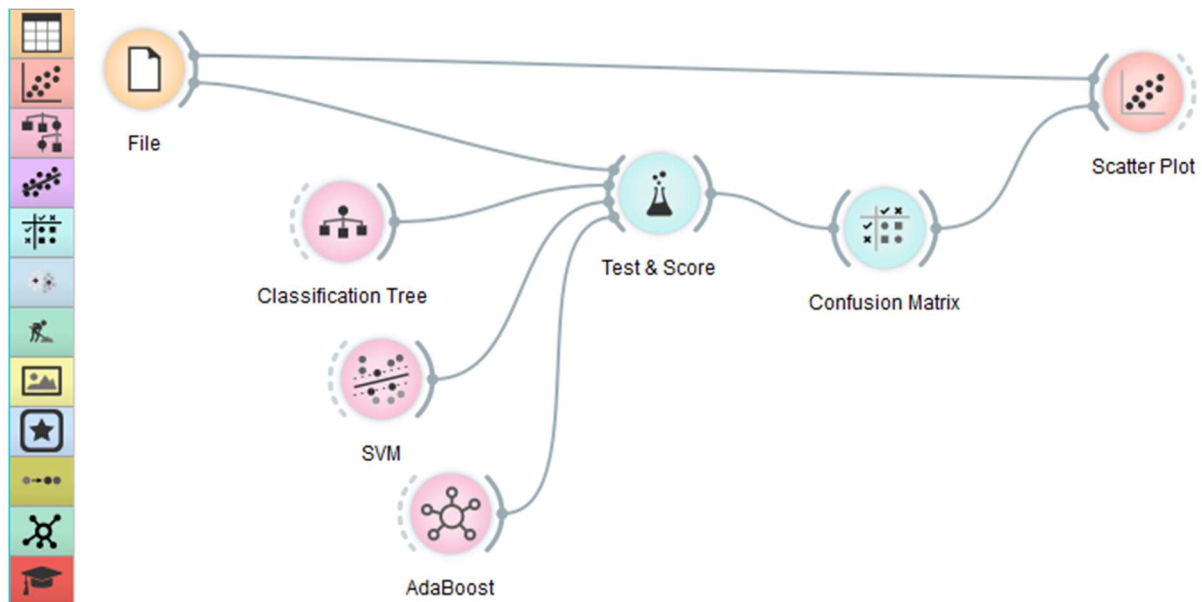


Figure 4-4: Orange Visual Programming (Demšar et al., 2013).

Orange can be found at their website: <https://orange.biolab.si/>

4.6 RapidMiner:

RapidMiner is subscription-based software platform that has been designed for analytics teams. The platform unites data preparation, machine learning models, and predictive model deployment into one visual workflow designer, similar to Figure 4-4. RapidMiner is divided into three sections: RapidMiner studio, RapidMiner server and RapidMiner Radoop. RapidMiner studio is designed for use by data scientist with normal sized datasets. RapidMiner server is designed for large teams of data scientists to collaborate on a shared model with a large dataset. RapidMiner Radoop is designed for extremely large datasets that would fall under 'Big Data'. RapidMiner studio falls within the scope of this study and will be examined in more detail. RapidMiner studio contains over 1500 machine learning algorithms and functions divided into 6 categories. The categories provided are as follows:

- **Data Access:** Provides capability to connect to any data source, in any format and at any scale. The module also provides power tools to access, load and extract information from unstructured text data in a number of different formats (pdf, web data, multimedia mining etc.)
- **Data Exploration:** Provides robust statistical overview capabilities to quickly explore the available data. Provides graphical display capabilities to provide intuitive understanding of the data. The module allows for the quick identification of missing values.
- **Data Blending/Preparation:** Provides capabilities to aggregate, filter, join, and sort data. The module provides tools for feature selection, creation and extraction to produce a good target dataset. The module offers advanced attribute weighting capabilities along with data quality measures.
- **Data Cleansing:** Provides powerful tools for cleaning of the data. The automatic identification and removal of duplicate entries, anomaly and outlier detection, normalisation and standardisation, and sophisticated data dimensionality reduction capabilities.
- **Modelling:** The module provides a large number of models for classification, regression and clustering modelling. Association mining, frequency item set mining and similarity computations are also available. The module seamlessly integrates with R and Python to enable custom scripts to be used. The module provides capabilities for advanced methods such as ensemble classifiers and parameter optimisation loops.
- **Validation:** This module provides all the required validation and evaluation measures and procedures that are required for every data mining type.

The visual workflow designer used by RapidMiner allows any user with an understanding of the data mining process but little programming knowledge to create and implement data mining applications. The RapidMiner studio module has been designed to increase productivity and aid rapid implementation of data mining/science projects. To this end, they have automated a large part of the process, allowing the user to quickly and easily analyse their data by selecting from a number of suggested processes and models, known as the AutoMiner. The AutoMiner has been designed to suggest the most appropriate machine learning models and input parameters and the accompanying pre-processing steps to achieve accurate results in a short time frame. The TubroPrep module was recently added to increase the speed of data preparation. RapidMiner can be found at their website:

<https://rapidminer.com/>

4.7 Evaluation of Data Mining Resources:

In order to aid the selection of a data mining resource, the resources presented in this chapter were evaluated against the requirements of the data mining examples provided in Chapter 3 (Lee *et al.*, 2011). The methods and algorithms required for each step within the data mining

process is summarised in Table 11-1 in Appendix 3. Presented alongside the requirements are the capabilities of each of the available resources.

Orange and RapidMiner: The two stand-alone software packages are able to fulfil nearly every requirement of the two data mining applications. This is expected as they are designed to provide every method in a pre-packaged format that is easy-to-use. However, they are not able to perform information extraction (see Section 3.3.1). Due to the limitations that visual interface places on the user, compared to the ‘freedom’ of the programming languages to adjust to any problem, it will not be possible to perform this function with either Orange or RapidMiner.

mlr: Mlr provides methods to perform nearly every function required by both applications. It does not provide methods for data imputation and information extraction. While mlr does not provide a function specifically for data imputation, it will still be possible to perform this function due to the implementation being in a programming environment. The advantage of a traditional programming environment is that the user can adapt the functionality provided by the environment and the data mining resource to perform a slightly altered function. Therefore, the user will ultimately be able to perform data imputation in the R environment where mlr can be used for nearly other requirement.

Scikit-learn: is focused on data mining and machine learning and not advanced text analysis. Therefore, Scikit-learn is able to perform all the required functions apart from information extraction. The wide range of functionality provided by Scikit-learn comes with the requirement that the user be proficient in programming in Python. This could be a hurdle for a potential data mining practitioner if they have been trained in a different programming language or have no programming experience.

NLTK: Due to NLTK being devoted to processing and analysing natural language, the resource fails to perform the majority of the required functionality. However, it is the only resource able to perform information extraction. The ‘wrapper’ functionality provided by NLTK enables other Python modules to utilise the processed text data it generates. This allows a data mining practitioner to utilise Scikit-learn and NLTK in conjunction to perform all the required functions.

Recommendations: Four of the resources provided are able to perform the vast majority of the functions required by the two applications presented in Chapter 3. The main shortcoming of these four methods is their inability to perform information extraction. Since information extraction is seen as a specialised form of data mining and is not as widely applicable, a limited number of data mining resources provide the capability to perform information extraction. While the prediction of the impact of on-site accident data mining example stated that information extraction is required (see Section 3.3.4), it might be the case that the on-site accident data could be obtained in a different way, thus removing the requirement for the information extraction.

If a data mining practitioner wishes to adopt an easy-to-use data mining resource and the practitioner does not require the ability to perform information extraction, it is recommended that they select either Orange or RapidMiner, depending on their budget. If the practitioner has experience programming in R, then MLR is recommended. If the practitioner has experience in Python and requires the ability to perform information extraction, then the combination of Scikit-learn and NLTK is recommended.

4.8 Conclusion:

The resources presented in Chapter 4 are well-rounded and user friendly. The list of resources presented is not an exhaustive list as there are many other resources that fulfil the same purpose. Due to the wide range of implementation resources and the varying levels of skill required to use these resources, it is clear that an inexperienced data mining practitioner will be able to select a resource that suits their level of expertise. The collaborative and helpful communities that surround these resources allow inexperienced practitioners to learn the environment and to learn more about data mining while creating an application. This rapid and collaborative learning will result in the practitioner creating more intricate and accurate models as they learn. A summary of Chapter 4 is provided in Appendix 3.

5 Data Mining Implementation Example:

5.1 Introduction:

Chapter 5 fulfils the fourth and final objective of the research by applying the data mining process defined in Chapter 2 and discussed in Chapter 3 to a real dataset. The demonstration of data mining will apply some of the resources presented in Chapter 4 to an internal project database dataset obtained from the Directorate of Construction and Maintenance in the Western Cape Governments' Department of Transport and Public Works.

A data mining application is implemented by following the data mining process discussed in Chapter 3 from the initial step, Goal Definition, through to the final step, Validation and Evaluation. The application aims to predict the number of employment opportunities a South African road construction project will create. Thereby reducing the uncertainty during the planning phase of the positive social impact of the project. This is an important consideration in a country with high unemployment rates, such as South Africa. The data available contains 60+ possible features, which are explored to determine if the goal is feasible. The data is pre-processed into a clean, normalised target dataset containing 4 features. The features are modelled using seven different classification algorithms and their accuracies are examined. The development of the application is described through the number of iterations it passed through.

A second data mining application is started by following the data mining process. The application aims to predict if a construction project will experience a cost overrun using a classification algorithm. The same dataset obtained for the first application is used in this application. After an initial investigation of the data it is concluded that the data does not contain the necessary information to accurately model the system. As the collection of new data is not possible, the application is abandoned.

The chapter presents a number of lessons learned during the data mining process that will prove useful for construction sector personnel who wish to implement data mining to aid project management.

5.2 Prediction of Employment Opportunities a Construction Project in South Africa will Create:

South Africa suffers from high rates of unemployment. The overall unemployment rate of the country in the first quarter of 2018 is 26.7% with youth unemployment being 38.2% (Khumalo, 2018). The South African government have committed to reducing unemployment by accelerating the creation of employment opportunities. Part of this strategy is to provide temporary employment opportunities to unskilled workers during construction projects funded by the government. The creation of a specified number of unskilled employment opportunities has become a project requirement imposed upon the contractor by the government. The number of employment opportunities a contractor needs to create is typically decided by government guidelines.

5.2.1 Goal Definition:

The goal defined in this step must reduce the uncertainty within the project regarding at least one of the five sustainable project criteria. The goal for this project is to provide estimates during the initiation or planning phases of a project of the number of temporary employment opportunities each project will generate by using a regression algorithm. This will help reduce the uncertainty regarding the positive social impact of the project during the initiation or planning phase of the project.

The information provided by the data mining application will provide the Directorate of Construction and Maintenance at the Western Cape department of Public Works with data-based estimates of the total number of employment opportunities their current and planned projects will create. The information provided by the application can be used on a project level when setting the target for the number of employment opportunities a project will be required to create. If the estimate and the guidelines vary on the number of employment opportunities the project will create, the department should consider both and make an informed decision.

5.2.2 Data Acquisition:

The data was acquired from the Construction and Maintenance Directorate of the Western Cape Department of Transport and Public Works' internal project databases. The information stored in these databases was obtained from the project managers on the projects, from the contractors on the project, from the engineers on the project and from other project documentation. Several data spreadsheets were collected and placed in the data warehouse. The spreadsheets contained information such as project details (project objective, tender details, initiation and completion dates, initial and final cost estimates etc.), project deliverable information, project subcontractor details and project temporary employment details.

5.2.3 Pre-Processing:

The pre-processing step is the most labour intensive in the data mining process if the application uses predefined data mining algorithms. The tools defined in Section 3.4 were used to in accordance to process depicted in Figure 3-3 to process the data stored in the data warehouse into a target dataset.

5.2.3.1 Goal Definition (Re-examination):

The goal of the project is to provide estimates of the number of temporary employment opportunities a project will create in the initiation or planning phase of the project. An initial investigation was done to determine if the goal of the application was feasible with the data collected and to determine the types of data pre-processing that will be required.

The feasibility of the goal was investigated by examining the data for correlations between the collected features and the number of employment opportunities created per project. The

strongest correlation found in the data was between the number of employment opportunities created and the final cost of the project, as can be seen in Figure 5-1:

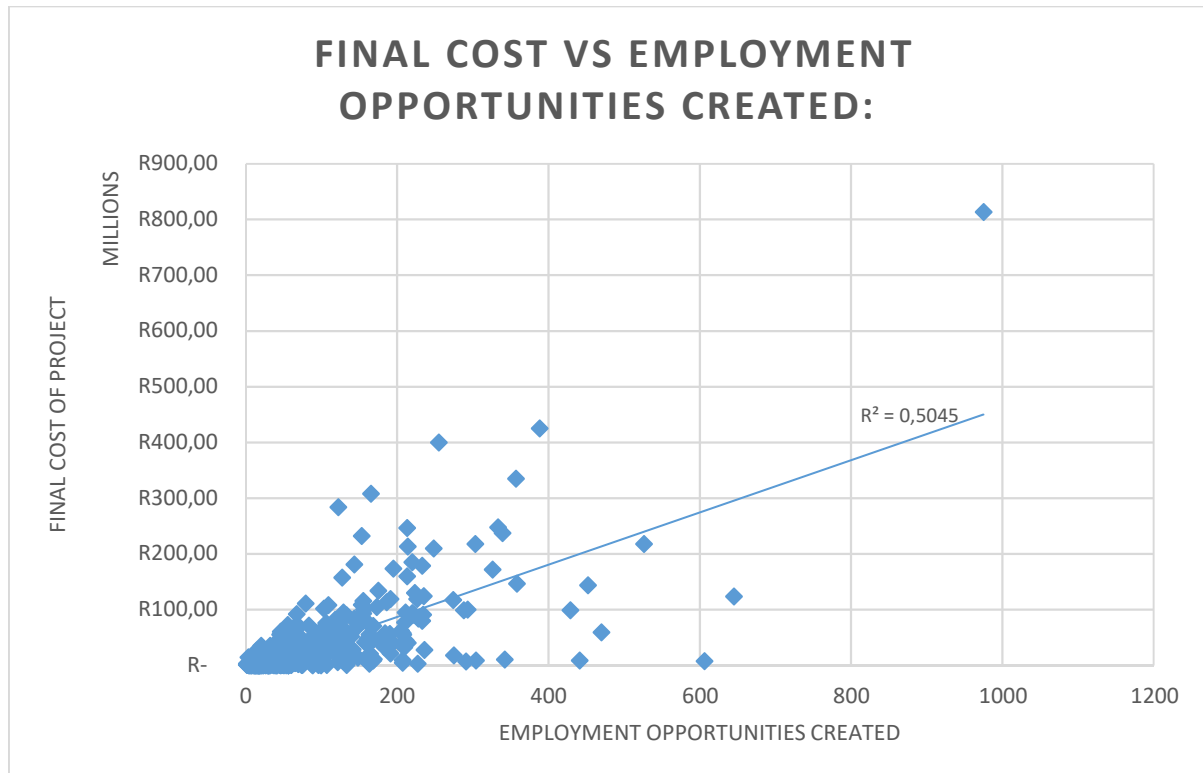


Figure 5-1: Final cost vs Employment Opportunities created

While the correlation between the number of employment opportunities created and the final cost of the project is the strongest correlation found in the data, the final cost of a project is not available during the initiation and planning phases of the project. Therefore, the final cost of the project cannot be included as a feature in the target dataset. The correlation between the cost of the project and the number of employment opportunities created is logical, as a more expensive project will typically require more labour for a longer period. The tender cost estimate provides similar information to the final project cost and is information that will be available near the end of the planning phase and can therefore be used as a feature in the target dataset. The correlation between the tender cost estimate and the employment opportunities created is shown in Figure 5-2:

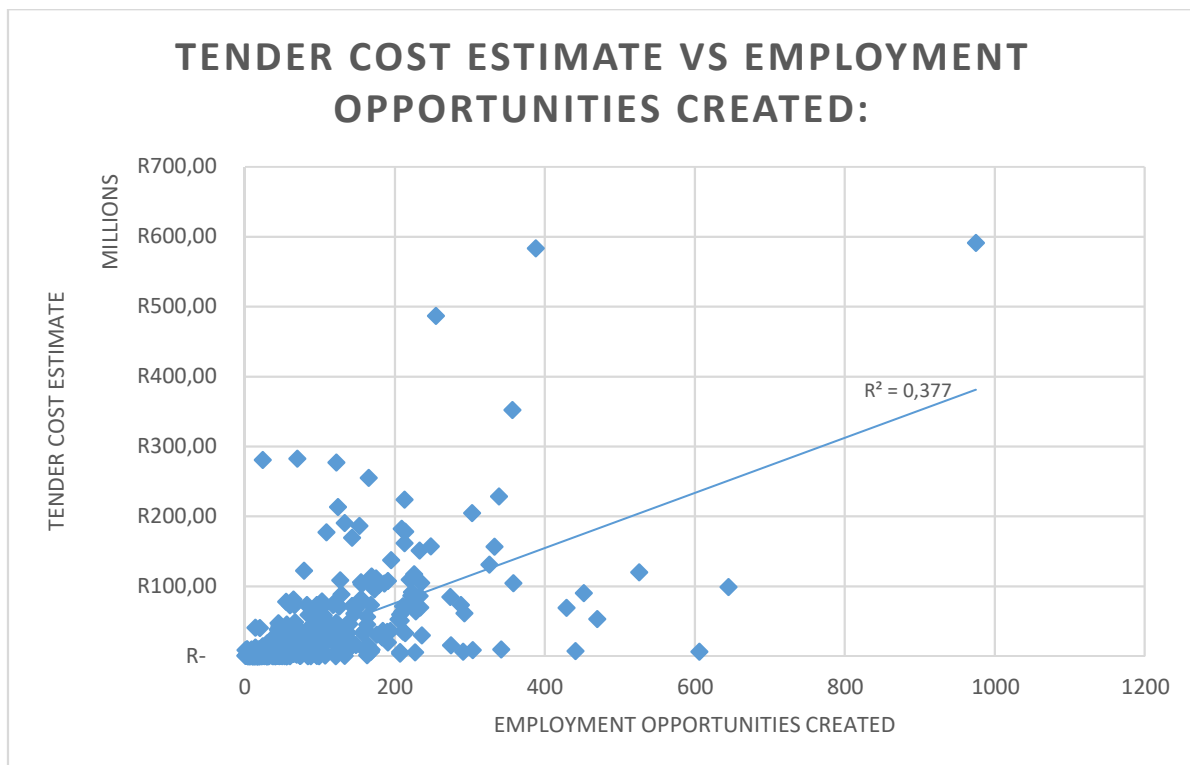


Figure 5-2: Tender cost estimate vs employment opportunities created

The correlation between the final project cost and the employment opportunities created ($R^2 = 0.5045$) is stronger than the correlation between the tender cost estimate and the employment opportunities created ($R^2 = 0.377$). This can be attributed to the fact that the number of employment opportunities created is information that is collected after the project has been completed. Therefore, the number of employment opportunities created will be influenced by scope changes or other major changes to the project, information that will affect the final cost of the project, but that the tender cost estimate cannot consider. To rectify this issue, it will be necessary to remove outliers where there were large increases or decreases in cost (usually indicative of scope change). It can also be seen that there are other outliers where the cost of the project was low, but the project created a large number of employment opportunities. These outliers must be examined.

Initially the goal was to be achieved by using a regression algorithm to provide exact estimates of the number of employment opportunities a construction project will create. After examining the relatively low coefficient of determination (R^2) between the tender cost estimates and the number of employment opportunities created, the type of data mining algorithm used was changed to a classification algorithm. This change requires the number of employment opportunities created to be binned into discrete numeric classes.

Despite some data quality issues and the change in data mining algorithm type, the correlation between the initial cost estimate and the employment opportunities created indicates that the goal of the project is feasible, and the investigation can be continued.

5.2.3.2 Feature Extraction and Selection, Data Cleaning, and Feature Transformation:

The data warehouse contained data on 755 road construction and rehabilitation projects and more than 60 possible features. Selecting the most influential features to include in the target dataset is important as irrelevant features can negatively impact the accuracy of the results while also dramatically increasing the computational power required to model the system. The vast majority of the features in the dataset can be ignored for this application. Features such as road number or age of the employee are irrelevant to determining how many employment opportunities will be created by a road construction project.

Those features that would be available during the initiation or planning phases and were selected to be further examined are listed below:

- District.
- Project objective.
- Tender cost estimate of the project.
- Tender construction duration estimate.
- Temporary employment opportunities created.

Each project had a district and a project objective assigned to it. The district feature was given as one of the seven municipal districts in the Western Cape. The project objective was given as one of thirteen possible main project objectives, such as 'routine surface maintenance', 'resealing/resurfacing', or 'bridge construction with surface'. These features were transformed from categorical text data to categorical numeric data. A sample of the transformation is shown in Table 12-1 in Appendix 4.

Each project had a tender cost estimate. These tender cost estimates range from hundreds of thousands of rands to hundreds of millions of rands. As discussed in Section 3.4.4, if the features within the dataset are of different scales, such as millions of rands and hundreds of employment opportunities, the features must be normalised or standardised to prevent the data mining algorithm from incorrectly modelling the feature of larger scale as being more influential. An example of the standardisation of the tender cost estimate for the road construction and maintenance projects used in the dataset is shown in Table 12-2 in Appendix 4.

The number of employment opportunities created and the tender construction duration estimate had a large number of missing values. Unfortunately, the number of projects that contained both the number of employment opportunities created and the tender construction duration estimate were too few and therefore the tender construction duration estimate was excluded from the target dataset.

Projects that showed a large difference between the tender cost estimate and the final project cost were omitted since the difference could indicate a change of scope. Such a scope change will impact the number of workers required and would therefore invalidate any predictions made based on the tender cost estimate.

The number of employment opportunities created was also examined, as there were several projects with a low tender cost estimate that had created a large number of employment opportunities. Many of these projects had duplicate worker entries that had to be removed. Other projects had provided short 4 to 5-week contracts to the employees. When the employee's contract was renewed the same employee would be assigned new employee ID. This resulted in several projects where a single person had many unique worker IDs. The number of employee IDs therefore did not represent the number of employment opportunities that were created. These duplicates were removed to provide an accurate count of the number of employment opportunities created that lasted the duration of the project.

Figure 5-3 shows the relationship between the number of employment opportunities created and the tender cost estimate after the outliers have been removed and the data has been cleaned. By comparing the R^2 - statistic of Figure 5-2 ($R^2 = 0.377$) and Figure 5-3 ($R^2 = 0.4274$), it can be seen that the removal of the outliers and cleaning the data has increased the correlation between the tender cost estimate (the most influential feature) and the number of employment opportunities created. This indicates that the pre-processing tools used were successful at removing noise from the target dataset.

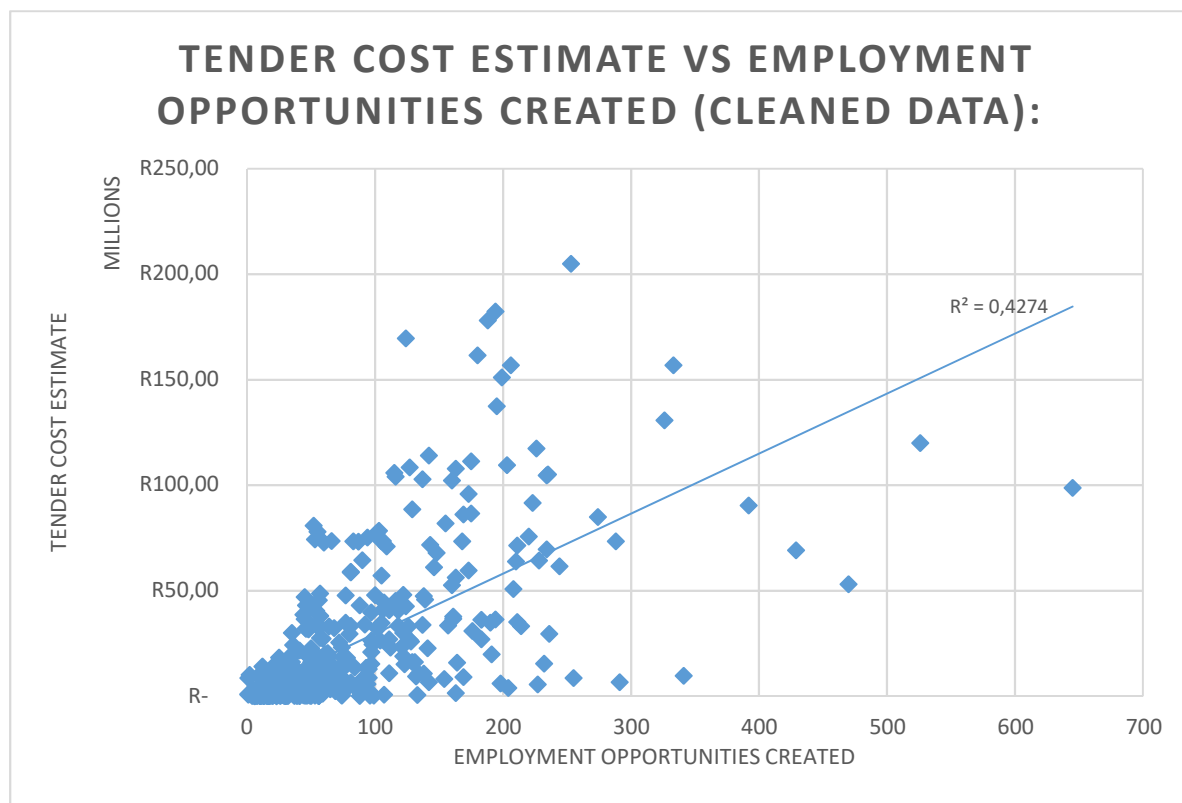


Figure 5-3: Tender cost estimate vs Employment Opportunities created (Cleaned data)

Despite the improved R^2 -statistic, classification algorithms were used for ease-of-use, ease-of-interpretation, and increased accuracy. In order to use a classification algorithm, Strugz' rule was used to determine the number of bins required. Binning is the process by which continuous variables (eg 1;2,4;5;7;8) are grouped into ranges and the variables that fall within those ranges are assigned the same bin value (eg 0 to 3 = 1 and 4 to 8 = 2). A sample of the

binning is shown in Table 12-3 in Appendix 4. The employment opportunities created feature was therefore transformed from a continuous numeric value to a discrete categorical value. Transforming the “employment opportunities created” feature from a continuous numeric value to a discrete categorical value via binning will result in the prediction precision being reduced as the number of employment opportunities created is no longer a single value but a bin that represents a range of values. The trade-off between prediction precision and prediction accuracy is acceptable in this case as the information gained from the application are estimates and will be used to supplement information from government guidelines.

At the end of the pre-processing process the target dataset contained the following normalised, cleaned and transformed features for 475 road construction or maintenance projects:

- District
- Project objective
- Tender cost estimate
- Number of employment opportunities created

Table 12-4 in Appendix 4 shows a sample of the target dataset used in this application.

5.2.4 Mining and Modelling:

The data mining algorithms were implemented using Python. The Scikit-learn Python module (see Section 4.2) was used along with the Numpy and Pandas modules. The Numpy module is a fundamental package for scientific computing that is required by both Scikit-learn and Pandas. The Pandas module is a data handling and data analysis module that was used to set up the target dataset. The machine learning models selected are listed below (all are discussed in Section 3.5.5):

1. K Nearest Neighbours classifier
2. Naïve Bayes (Gaussian) classifier
3. Naïve Bayes (Multinomial) classifier
4. Support Vector Machine classifier
5. Decision Tree classifier
6. Neural Network classifier

These 6 classifiers were chosen to demonstrate the variety of classifiers available in the Scikit-learn module. The same training and testing sets were used for each classifier, so their accuracies could be compared. The implementation of the application in Python is shown in Figure 12-4 in Appendix 4.

5.2.5 Validation and Evaluation:

The validation process used for the classifier was a 10-time hold out testing process. The target dataset was randomly split into the training dataset (80%) and the testing dataset (20%). The classifier was trained and tested on the two sets and the accuracy recorded. This

process was repeated 10 times, with the training and testing dataset being randomly assigned 80% and 20% of the target dataset respectively. The total accuracy reported below is the average accuracy for each classifier over the 10 runs.

Table 5-1: Employment Opportunities Creation Prediction Accuracies:

Algorithm:	Total Accuracy:
K Nearest Neighbours Classifier:	60.7%
Naïve Bayes (Gaussian) Classifier:	62.3%
Naïve Bayes (Multinomial) Classifier:	55.5%
Support Vector Machine Classifier:	58.3%
Decision Tree Classifier:	53.2%
Neural Network Classifier:	55.9%

As can be seen above the classifier that performs the best over the ten runs is the Naïve Bayes (Gaussian) Classifier. Decoding the exact reasons for certain classifiers performing well while others perform poorly is a complex and difficult task that is influenced by many factors. Experts in the field of machine learning and statistics will be able to select the most suitable algorithm for a given problem from the outset, but the ease of using the different models provided by Scikit-learn allows the user to implement a number of algorithms and then to select the most accurate algorithm. This approach allows a novice data mining practitioner to select the best model and to achieve good prediction results.

The overall accuracy obtained from the best-performing classification algorithm is 62,3%. Whether a prediction accuracy of 62.3% should be accepted and the model put into use is debatable. When predicting classes in an application with very well-established causal links or pre-existing methods (such as the medical field), the required accuracy is very high. By contrast, the acceptable accuracy when predicting classes in an application with a large number of influencing factors and some randomness is less. In case study 4, discussed in Section 2.3.4, Williams and Gong (2014) attempted to predict the cost overrun of construction project, but only managed to obtain an overall average of 44% for 3 possible classes. This application, however, was able to achieve a prediction accuracy of 62.3% for 12 possible classes. The complexity of modelling construction projects to predict cost overruns is admittedly much greater than modelling construction projects to predict the number of employment opportunities created.

An aspect that must be considered before accepting or rejecting a data mining mode is the mode of failure of the prediction algorithms. In this application, the Naïve Bayes (Gaussian) algorithm predicts the correct class with 62.3% accuracy. When the algorithm predicts the class incorrectly, it often chooses the class one above or one below the correct class. This could still be useful as the information provided by these incorrect classifications still provides a ‘ball-park’ figure that can be used in broad, departmental wide planning.

Ultimately, as stated in Section 3.6, whether the model should be accepted or rejected based on its prediction accuracy is up to the investigator, the goals of the application, and the specific requirements of the department.

5.2.6 Iterations:

The process presented above is the final process that was used for this application. However, as shown in Figure 2-6 (data mining process figure), the data mining process is both linear and iterative. Listed below are the iterations of the data mining process and the changes made with each iteration.

1. Switch from Regression modelling to Classification modelling. The initial goal of the application was to model the number of employment opportunities created using a regression algorithm. However, the relatively low R^2 -statistic obtained during the initial data exploration indicated that regression modelling would not be appropriate. Therefore, the goal was adjusted to use classification modelling.
2. Increase in employment opportunities bin size. Initially the number of employment opportunities were binned into 25 bins. The model was trained and tested on the target dataset, but the results obtained were very poor. The number of bins was subsequently reduced which resulted in increased prediction accuracies.
3. Investigation of outliers. During the initial iterations the data was cleaned and processed but no outliers were removed. After the number of employment opportunities bins had been increased, the outliers were investigated and removed, as described earlier.
4. Addition of models. The three previous iterations were done using only a Decision Tree algorithm. 5 additional models were added to enable the best model to be selected.
5. Refining of process and increased evaluation rigor. Once the 7 models were working, the implementation was refined with small changes being made to the input parameters and the normalising process. The 10 times hold out testing-training process described earlier was introduced to ensure that the results obtained were valid.

5.3 Prediction of Project Cost Overruns:

Cost overruns on construction projects are a global problem. The highly competitive nature of the construction industry and the complexity of construction projects means the factors that can lead to a project experiencing a cost overrun are numerous (Mukuka, Aigbavboa and Thwala, 2015; Senouci, Ismail and Eldin, 2016). Cost and time overruns are a major challenge for developing countries where many projects are focused on infrastructure and designed to provide the population with basic services (Niazi and Painting, 2017). These cost overruns limit the government's ability to adequately meet its development goals and can lead to a slow-down in project roll-out.

5.3.1 Goal Definition:

The ability to predict if a project will experience a cost overrun will a) significantly increase the ability of project managers to recognise and respond to issues before they materialise and b) allow government and private client bodies to allocate resources to projects that are likely to experience overruns.

The goal of this application is to predict, by use of a classification algorithm, whether a project will experience no cost overrun, a moderate cost overrun, or a significant cost overrun. The goal of this application is to significantly reduce the uncertainty, during the planning phase, regarding the cost of the project.

5.3.2 Data Acquisition:

The data used for this application is the same base datasets provided by the Western Cape Government's Department of Transportation and Public Works that was used in the previous application (see Section 5.2.2).

5.3.3 Pre-processing:

The pre-processing step started with the re-examination of the goals of the application in light of the collected data. The goal of the application is to predict whether a project will have no cost overrun, a moderate overrun or a significant overrun, by using a classification algorithm. A preliminary investigation into the available data was conducted to determine the feasibility of the application.

The initial investigation into the data indicated that the data collected was not suitable for this application. This is due to the very poor correlation between the available features (number of variation orders, initial duration of the project etc.) and the output class (cost increase). Figure 5-4 shows the relationship between the percentage cost increase and the number of variation orders a project experienced. The correlation between these features were the highest within the dataset but was still very poor ($R^2 = 0.0425$).

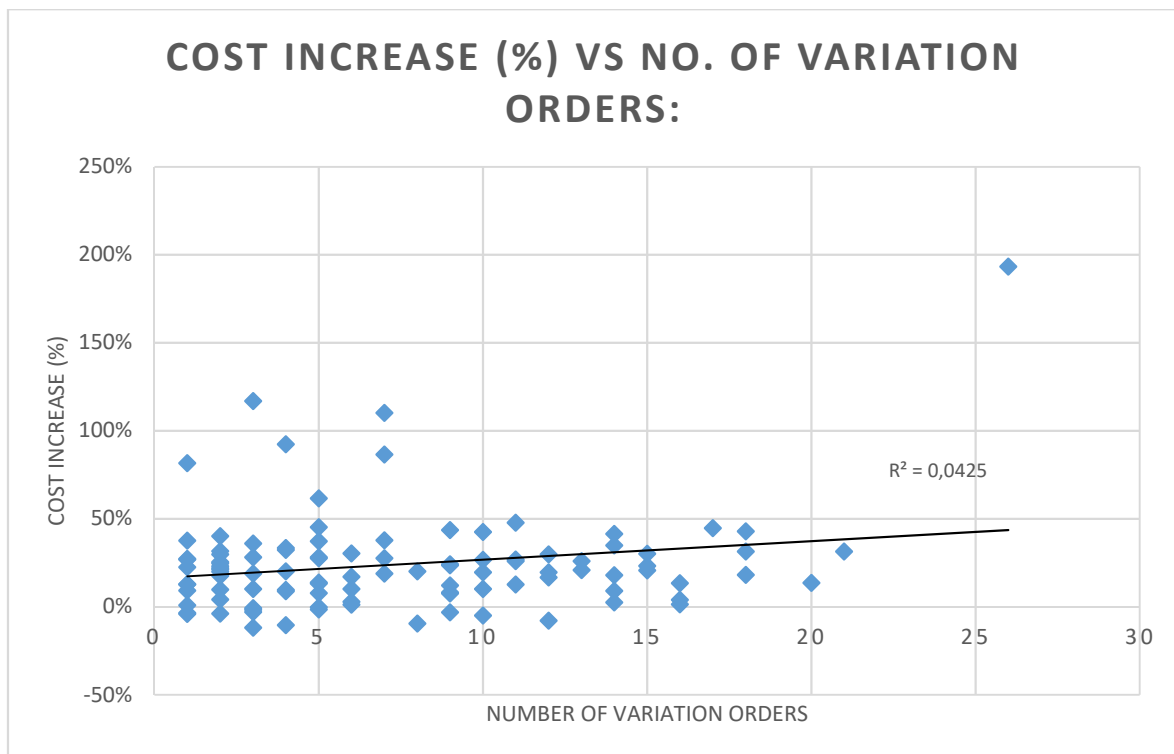


Figure 5-4: Cost (%) vs Number of Variation Orders

In the data mining application presented in Section 5.2, the initial correlation between the most informative feature and the class label feature wasn't poor ($R^2 = 0.377$). It was increased (to $R^2 = 0.4274$) after the pre-processing step. Ultimately an accuracy of 62.3% was achieved by the best data mining model. While the 62.3% accuracy obtained shows that data mining algorithms can create far better models than a linear regression model, it implies that a useable model could not be created from features that are as uncorrelated as those collected for this application. The application was therefore deemed unfeasible and shelved until more appropriate data could be collected.

5.3.4 Discussion of Application Failure:

The application was deemed unfeasible due to the poor internal relationships within the data collected. This does not imply that such an application is impossible, as otherwise proven by (Lee *et al.*, 2011) who achieved good prediction results. This merely indicates that the application is not feasible with the data currently available. This could be remedied by the collection of more appropriate data. Lee *et al.* (2011) and other similar applications that achieved success with a similar goal can be examined to determine what data is required to successfully implement a cost overrun prediction application.

The data mining model used by Lee *et al.* (2011) was a Support Vector Machine (SVM) configured for regression modelling. This algorithm type is widely available in the data mining resources presented in Chapter 4. Therefore, if the required data was collected and an SVM regression model was used to model the data, there should be no reason why the results from Lee *et al.* (2011) could not be replicated.

Once the required data has been identified, a system that collects this data from ongoing and completed projects can be implemented. Unfortunately, this research was not able to collect the required data and as such the application could not be completed.

5.4 Lessons Learned from Applying the Data Mining Process:

This section aims to provide a list of lessons and possible additions to the data mining process based on the experience gained from applying the data mining process to the project data provided by the Western Cape Government Department of Transportation and Public Works.

Initial Investigation of the Data: Conducting an initial investigation of the data to determine if the goals of the application are feasible is mentioned in Section 3.2. While this investigation is extremely application specific and could not be discussed further in Section 3.2, it is very important. A good preliminary investigation of the data will inform the investigator not only if the application goals are feasible, but also of the problems the dataset might contain. Data quality issues such as missing values, outliers, inconsistencies in data type or scale, duplicates and other should be searched for to determine the techniques that will need to be employed to produce a good target dataset.

Specifying the Data Mining Types During Goal Definition: The goal definition described in Section 3.2 focuses mainly on the need for the goal to reduce uncertainty within the project with regards to one of the sustainable project management success criteria. The section mentions that the data mining model type should be defined but does not stress how important it is. The type of data mining the application uses will determine how it is implemented. Classification can assign class labels, regression can assign a continuous variable, clustering can group similar objects and outlier detection can identify points that are irregular. These different data mining types will be employed in different ways and will significantly influence every aspect of the application and how its results are used.

Ensuring the Data Contains the Necessary Information to Fulfil the Goal: The definition of an application goal should be conducted when the data environment is known. Defining a goal with a wide and ambitious scope, as was done in Section 5.3, without first being familiar with the data that is available will inevitably lead to the scope of the goal being reduced or the application being abandoned.

The data collected for the data mining implementations was initially collected for the purpose of predicting the number of employment opportunities a construction project will create. As such, it contained the required information to fulfil that goal (such as the objective of the project, the cost, the number of workers, the initial duration etc.). This information is not sufficient to allow for the fulfilment of a more ambitious goal such as predicting cost overruns on a project. Lee *et al* were able to accomplish this goal by obtaining detailed project planning and scoping information. In contrast, little scope information was included in the data set used for the two data mining implementations.

To accurately model a system all the factors that influence the system must be accounted for in the data collected. The omission of scope information from the dataset of the second implementation example resulted in the application being shelved.

Data Availability During the Project Lifecycle: The data collected, and features selected to model a problem should be available at the time in the project lifecycle when the application will be applied. In the application to predict the number of employment opportunities (Section 6.2) the feature with the highest correlation to the number of employment opportunities created is the final project cost. The final project cost cannot be used to predict the number of employment opportunities created during the planning phase as the final project cost is not available during the planning phase of the project. Therefore it is important to ensure that the data used for the application is available when the application will be implemented.

Familiarity with the Chosen Data Mining Resource: Familiarity with the data mining environment is important. Whether the environment chosen is one presented in Chapter 4 or a different environment, familiarity with the data mining environment can be achieved by examining the environment documentation. Special attention should be paid to any method, technique or model that could be used in an application.

This is due to small differences between method requirements that could exist within the environment. The assumption cannot be made that all the methods will be able to utilise the same dataset as small differences between input data requirements for different models could exist. For example, a classification model might not require the data to be normalised, as it automatically performs this function, whereas another model might require the data to be normalised. Therefore, the data will need to be normalised if both models are to be used.

It is not essential to understand exactly how the data mining model or technique works, but rather to understand its data input requirements and possible model limitations.

Utilising Several Data Mining Algorithms: The ability to easily utilise several data mining algorithms is a significant advantage of the data mining environments presented in Chapter 4. If the model limitations and requirements are understood, it is possible to develop a target dataset that is suitable for all the models intended to be used.

Using several models and suitable validation criteria grants the distinct advantage of foregoing the need determine which model will be the most appropriate at the outset of the investigation. This is particularly significant for novice data mining practitioners who may not understand the implementation and limitations of every model.

The models can be trained and evaluated to determine which model is the most suitable for the application. That model may then selected and put to use.

Iterative Nature of the Data Mining Process: The data mining applications presented often depict the process as a linear process. Some applications mentioned that the process is iterative, but few describe the number of iterations the applications underwent before the final result was obtained. The iterative nature of the data mining process should be seen as an advantage as it provides novice data mining practitioners the ability to fine-tune their application, address any issues that may have arisen and to continually learn and improve their data mining knowledge. However, note must be taken of the changes made to the data as removing or overworking the data can lead to a target dataset that is not representative of the system being modelled.

Accuracy and Usability of the Model: Whether the accuracy obtained by the application is acceptable or not will depend on the goal of the application. The model should aim to reduce the uncertainty with a project at a specific phase. If a method for providing the required information does not currently exist and the inherent uncertainty with the project is high when the model is being applied, then the model accuracy required to use the predictions will not be very high. If the inherent uncertainty is low and accurate methods do exist to provide the information, then the required accuracy for the data mining application will be much higher.

The usability of the model must also be considered. If the model is designed for a once off application, then the model is not required to be extremely streamlined and fast. If the model is designed to be used a number of times over a period of time, then the model will need to be able to handle additional data and many predictions without the investigator having to manually initialise the model or compile the target dataset.

Additional Data: After the data acquisition step, if the available data is insufficient to model the problem, it may be necessary to acquire additional data. Additional data may be collected through the creation of data collection and recording systems within an organisation if the application has the backing of senior management. This can be done at government entities and private client bodies by requiring the contractor and engineering team to submit a document or spreadsheet with predefined data fields filled out. This will increase the available data and reduce pre-processing requirements.

5.5 Conclusion:

Chapter 5 fulfilled the fourth objective of the research. A data mining application was developed using the process defined and discussed in Chapter 2 and 3 and using (data mining) resources presented in Chapter 4. The data mining implementation example presented within this chapter demonstrates that data mining can be used in the construction environment to improve the management of a project. However, simply following the data mining process defined in Figure 2-6 does not guarantee that a successful application will be created. A number of valuable lessons were learned during the implementation of the data mining examples. These lessons must be taken into account before implementing a data mining application as it will reduce the possibility of producing an application with lower than acceptable prediction accuracies.

6 Conclusion:

The use of data mining within the construction sector to improve project management has been limited. This research therefore aimed to determine a process for applying data mining to project management in the construction sector and to demonstrate the process with a data mining application on a real dataset (Section 1.2).

6.1 Synthesis and Discussion of the Data Mining Process:

Aim 1 of this research was to synthesis and discuss a data mining process specifically designed for application to construction projects to facilitate project management. This aim was divided in to two objectives: 1) synthesis of the data mining process and 2) discussion of the data mining process. These two objectives were met in Chapter 2 and Chapter 3 respectively

Eight case studies were examined to determine if the construction sector is suitable for data mining to improve project management and to synthesise a data mining process. The research investigated 5 case studies of data mining in the construction sector and 3 case studies of data mining in the software development sector where it was used to aid project management.

The similarity between software development projects, both in terms of the phases of each project and the uncertainty within each phase, indicated that the environment is suitable for data mining. The significant amount of data that is stored in project databases and within project documentation indicated that the construction sector is suited for the application of data mining to improve project management.

Three case studies explicitly defined the data mining process they adhered to. These processes formed the foundation of the data mining process synthesised in Chapter 2. Additional information was gathered from examining the other case studies and was combined with knowledge about construction projects to inductively synthesise the data mining process.

The process synthesised is presented below along with some of the key discussion points presented in Chapter 3, where the process was discussed extensively. The process was described with the goal of informing construction industry personnel, with little knowledge of the possibilities and the requirements of data mining, of the width of the field of data mining.

- **Step 1: Goal Definition:** The goal must reduce the uncertainty/risk exposure during the project, preferably during the initial stages, in terms of one of the five sustainable project management success criteria. The goal must define how the information will be used and what type of data mining will be used. Reducing the uncertainty by providing valuable information/estimates will aid project management.
- **Step 2: Data Acquisition:** In the construction sector the main sources of project data is project databases and project documentation. The data must be collected for as

many projects as possible while remaining cognisant of the goal of the application and the information that will be required to achieve the goal.

- **Step 3: Pre-Processing:** The collected data must be prepared into a target dataset.
 - a. Goal Re-examination: The feasibility of the goals of the application must be re-examined in light of the data collected. Any required changes in scope or execution methods must be made.
 - b. Feature Extraction and Selection, Data Cleaning, Data Reduction and Feature Transformation: The data collected must be processed, cleaned and outliers must be removed. The most informative features must be selected, extracted and transformed if necessary. These tools are critical to creating a suitable target dataset.

- **Step 4: Mining and Modelling:**
 - a. Similarity and Distance: A method for determining the similarity between data points. These methods are rarely used on their own, but often form the basis of other machine learning methods.
 - b. Association Pattern Mining: A methods for determining the relationships between features in a dataset using their frequency of occurring together in data entries. These methods were invented and are extensively used in the field of supermarket shopping behaviour analysis.
 - c. Cluster Analysis: These methods provide the capability to examine a dataset and to group the data into smaller groups of very similar data. This can be a very useful form of data summarisation.
 - d. Outlier Analysis: Outlier analysis is a complementary process to cluster analysis. Outlier detection algorithms cluster data points into similar groups but then also look for data points that do not conform to these internal data patterns. These non-conforming data points are labelled as outliers. Outlier detection is often used in data pre-processing.
 - e. Classification and Regression: Classification and Regression models both learn the internal data patterns from a labelled training dataset. A classification model utilises the learned internal data patterns to assign a class label to unlabelled data. A regression model utilises the learned patterns to assign a continuous numeric value to the data. These models are the most common data mining models and are powerful due to their ability to learn from example.
 - f. Text Mining: The mining of text data requires a number of specialised pre-processing steps, such as stop word removal and lemmatising, along with modified algorithms that take into account the special nature of the processed

data. Once the text has been processed into a vector representation, it is possible to apply clustering and classification models.

- **Step 5: Validation and Evaluation:** The output of the data mining models must be validated, and its accuracy determined.
 - a. Cluster Validation: The validity of clustering algorithm output is examined in one of three ways. Internal validation criteria measure the algorithm's effectiveness based on characteristics of the data. External validation criteria measure the algorithm's effectiveness by using synthetic external data sources. Relative validation criteria compare the results to other results by the same algorithm or results from other algorithms.
 - b. Outlier Validation: The validity of outliers can be tested using internal validation criteria, but this is hardly used due to the internal measuring bias. External validation criteria are used most often, with synthetic datasets or ROC curves being employed.
 - c. Classification and Regression Evaluation: The evaluation of classification and regression models is done using specific performance measures applied during an evaluation procedure. The performance measures differ for classification and regression, but the evaluation procedures are the same.

The data mining process is both a linear and an iterative process. The iterative improvement is a strength of the process because after completing a step the investigator is able to return to a previous step to make alterations to the application. These alterations must aim to increase the accuracy or usability of the application while achieving the set goal.

6.2 Data Mining Resources:

Aim 2 of this research had two objectives. The first of these objectives was to discuss the available data mining resources that could enable construction sector personnel to apply data mining. This objective was met in Chapter 4.

Data mining is a combination of statistics and computer science. Therefore, the implementation of data mining is typically done in a programming environment. Without a strong background in computer science and statistics the implementation of data mining from first principles in a programming environment is a daunting task. A number of data mining resources have been developed to significantly streamline the implementation of data mining by predefining the required tools, algorithms and techniques. The resources are available to the general public. These data mining toolkits are either based in a programming language, such as Python or R, or are stand-alone software packages.

5 data mining resources were selected and presented in Chapter 4. The resources presented were chosen due to their user-friendliness and completeness as data mining packages. The

completeness of the package was determined by examining the features of the resources for the following methods:

- Data pre-processing capabilities.
- Data modelling capabilities such as clustering, classification, and regression.
- Model validation techniques for all the model types contained in the package.
- Extensive documentation containing implementation guides and explanations of the methods contained.
- Text mining capabilities were not required but were counted as a bonus.

The resources presented and discussed were:

- **Scikit-learn:** Scikit-learn is a free data mining and machine learning module for Python. The module contains methods to pre-process data, conduct clustering analysis, conduct outlier analysis, conduct regression and classification, conduct text mining, and to perform the necessary validation processes for each of these. The module has a detailed implementation guide with thorough explanations of all the methods contained.
- **NLTK:** A free text mining Python module. NLTK has an extensive library of text processing methods. The module contains some classification and clustering capabilities, with the accompanying validation techniques, but provides ‘wrappers’ to allow the user to pass the processed text data to more sophisticated data mining packages. A free textbook accompanies the module with detailed explanations of text mining within the module.
- **Machine Learning with R (mlr):** A free machine learning and data mining package for R. The module provides a framework for clustering, classification, regression, and data pre-processing methods. The package provides the capability to evaluate all the methods as well as to benchmark them to determine the most computationally efficient method.
- **Orange:** A free stand-alone software package for data mining, text mining and information extraction. Orange uses a visual programming interface while providing the user with a large number of data processing, data modelling, and data validation techniques. The module contains several add-ons that enhance its capabilities. All the methods contained in the package are well documented with detailed implementation guides.
- **RapidMiner:** A subscription-based machine learning, data mining and text mining stand-alone software package. RapidMiner provides over 1500 methods for every step of the data mining process. A visual programming environment and well-documented methods increase the user-friendly aspect of the software package.

The list of resources presented are all suitable for a novice data mining practitioner from the construction sector. The resource(s) chosen for a specific investigation will depend on the

investigator's programming skills, their specific modelling needs, and their budget. These resources significantly lower the bar to enter the field of data mining, while still providing rigorous and highly customisable methods, for those willing to invest time and effort into a valuable new skill.

6.3 Data Mining Demonstration:

Aim 2 of this research contained two objectives. The second of these objectives was to create a data mining application by following the defined data mining process and using one of the available data mining resources. This objective was achieved in Chapter 5.

The demonstration of the data mining process aimed to provide an example of data mining in the construction sector with the specific aim to improve project management. The implementation of the data mining process showed how to apply the process but also serve as a check to ensure that no steps or information was omitted.

The data mining process was applied to a data set provided by the Western Cape Government's Department of Transport and Public Works. The data set was gathered from their internal project database and provided information on 755 road construction and rehabilitation projects and contained more than 60 possible features. A summary of the data mining process is presented below:

- 1. Goal Definition:** The application must provide an estimate, at the planning phase, of the number of unskilled employment opportunities a construction project will create. This information will help to reduce the uncertainty with regards to the positive social impact the project will have. The need for the information is due to the Government attempting to combat unemployment by providing work opportunities for unskilled labour in government funded construction projects.
- 2. Data Acquisition:** The data was collected from internal project databases from the Western Cape Government.
- 3. Data Pre-processing:** The goal of the project was re-examined after an initial investigation into the data collected. The goal was deemed to be feasible if a classification algorithm was used. The data was prepared by selecting 4 features, by cleaning the data of outliers and anomalies and by normalising and transforming the data. A well-prepared target dataset was produced.
- 4. Data Mining and Modelling:** The investigation implemented the application in Python by using the Scikit-learn module. 6 data mining algorithms were used.
- 5. Validation and Evaluation:** The data mining models were evaluated using a 10-times hold out process. The Naïve Bayes (Gaussian) Classification was the most accurate of the 6 models tested.

- 6. Iterations:** The data mining process was applied with 5 iterations. These included modifying the goal of the application, 3 additional pre-processing steps, and the addition of more models and a rigorous testing procedure.

A second data mining application was started with the aim of predicting if a construction project will experience a cost overrun. The application utilised the same dataset as was used in the first application. After an investigation in the data it was determined that the data did not contain the required information to accurately model the problem. The application was ultimately abandoned.

The application of the process provided valuable insight into the completeness of the data mining process as described in Chapter 3. Nine separate lessons were learned during the application of the process. Due to the data mining process being defined in broad manner to include all investigation into project management in the construction sector, some details were not stressed for some application specific scenarios. This is understandable as a broadly defined process will not be able to provide information and advice on how to handle every scenario. Rather, the process aims to serve as a guide to the type of methods that might be used to achieve the specific aims of each step in the data mining process, which it achieves.

7 Recommendations for Further Research

Chapter 7 provides a number of possible future avenues of research around data mining to improve project management in the construction sector.

Application of Data Mining to Real World Project Management: The research defined a data mining process and demonstrated the process on a real dataset. The research aimed to prove that it is possible and that it can be valuable. However, the application was limited in scope and did not achieve very high accuracies. A good opportunity exists for research into a more ambitious data mining application into improving project management by using data mining. The investigation could pair with a large construction industry partner to gather and collect a very large and varied dataset. There are several possible opportunities for research within this domain. Some of these opportunities are:

- Determine and demonstrate the applicability of large-scale information extraction from project documentation.
- Implement a data mining application for each of the three main stakeholders in a project (Client, Engineer and Contractor).
- Determine the different levels of data availability at each of the three main stakeholders and how data mining could be used to improve the management of projects from their perspective.
- Develop a number of easily implementable data mining applications for the mass market that can be used by any organisation.

These research topics will accelerate the adoption of data mining to aid project management in the construction sector.

Investigation into how Data Mining can Improve Project Management at Every Stage of the Project Lifecycle: The reduction of uncertainties within the project in terms of one of the sustainable project success criteria is the criteria adopted for the research on how to improve project management. This is a broad approach that can benefit from further investigation. Research into how data mining can be specifically applied at every stage of a project will help to establish the use of data mining for project management within the construction sector.

Training Required for Civil Engineers/Project Managers to Apply Data Mining: The research presents a number of predefined data mining resources and a robust data mining process that can be used by relative data mining novices. The adoption of data mining within the construction sector will increase rapidly if adequate and specific training opportunities are provided. Research should be conducted into the type of training that will be required to bring civil engineers, project managers and other trained construction industry personnel into the data mining field. The research can focus on smaller scale data mining applications with pre-

defined methods, such as was presented in this research, and on large scale complex data mining applications and the different levels of training required.

Study of Data Availability within BIM Environments: The research mentions building information models as a possible source of data for data mining. This is due to the central and collaborative nature of BIM and the ability to attach data and documentation to the model. Therefore, all the data about a project will be stored in a central location in a digital format. If BIM is shown to contain enough information, data mining could be included as functionality within BIM software.

8 References:

- Aggarwal, C. C. (2015) *Data Mining: The Textbook*, Springer International Publishing. doi: 10.1007/978-3-319-14142-8.
- Akintoye, A. S. and MacLeod, M. J. (1997) 'Risk analysis and management in construction', *International Journal of Project Management*, 15(1), pp. 31–38. doi: 10.1016/S0263-7863(96)00035-X.
- Alsabti, K., Ranka, S. and Singh, V. (1997) 'An efficient k-means clustering algorithm', *Electrical Engineering and Computer Science*. Available at: <https://surface.syr.edu/eecs/43>.
- Alsubaey, M., Asadi, A. and Makatsoris, H. (2015) 'A Naïve Bayes approach for EWS detection by text mining of unstructured data: A construction project case', *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, pp. 164–168. doi: 10.1109/IntelliSys.2015.7361140.
- Alvarez-Macias, J. L., Mata-Vazquez, J. and Riquelme-Santos, J. C. (2004) 'Data mining for the management of software development process', *International Journal of Software Engineering and Knowledge Engineering*, 14(6), pp. 665–695. Available at: <http://dx.doi.org/10.1142/S0218194004001841>.
- Archibald, R. D., Di Filippo, I. and Di Filippo, D. (2012) 'The Six-Phase Comprehensive Project Life Cycle Model Including the Project Incubation / Feasibility Phase and the Post-Project Evaluation Phase', *PM World Journal*, (2012), pp. 1–33.
- Balsera, J. V. *et al.* (2012) 'Data Mining Applied to the Improvement of Project Management', in *Data Mining Applications in Engineering and Medicine*. C.A: Intech. doi: 10.5772/46830.
- Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python*. 1st edn. Sebastopol, CA: O'Reilly Media Ince.
- Bischl, B. *et al.* (2016) 'mlr: Machine Learning in R', *Journal of Machine Learning Research*, 17, pp. 1–5.
- Borek, A. *et al.* (2013) 'Data and Information Assets', in *Total Information Risk Management: Maximising the Value of Data and Information Assets*. Elsevier Inc, pp. 3–22.
- Caccamese, A. and Bragantini, D. (2012) 'Beyond the Iron Triangle: Year Zero', in *PMI® Global Congress - EMEA, Marsailles, France*. Newtown Square, PA: Project Management Institute. Available at: <https://www.pmi.org/learning/library/beyond-iron-triangle-year-zero-6381>.
- Chaovalitwongse, W. A. *et al.* (2012) 'Data Mining Framework to Optimize the Bid Selection Policy for Competitively Bid Highway Construction Projects', *Journal of Construction Engineering and Management*, 138(2), pp. 277–286. doi: 10.1061/(ASCE)CO.1943-7862.0000386.
- Chinchor, N., Hirschman, L. and Lewis, D. D. (1993) 'Evaluating message understanding systems: An analysis of the third Message Understanding Conference (MUC-3)',

Computational Linguistics, 19(3), pp. 409–449.

Construction Industry Council (2013) ‘Building Information Model (Bim) Protocol’, p. 15. Available at: <http://cic.org.uk/download.php?f=the-bim-protocol.pdf>.

Demšar, J. *et al.* (2013) ‘Orange: Data Mining Toolbox in Python’, *Journal of Machine Learning Research*, 14, p. 23492353.

Ebbesen, J. B. and Hope, A. (2013) ‘Re-imagining the Iron Triangle: Embedding Sustainability into Project Constraints’, *PM World Journal*, 2(Iii), pp. 1–13. doi: 10.1146/annurev.nutr.26.061505.111236.

Fan, H., Abourizk, S. and Kim, H. (2008) ‘Assessing Residual Value of Heavy Construction Equipment Using Predictive Data Mining Model’, 22(3), pp. 181–191.

Fang, D. *et al.* (2004) ‘Risks in Chinese Construction Market—Contractors’ Perspective’, *Journal of Construction Engineering and Management*, 130(6), pp. 853–861. doi: 10.1061/(ASCE)0733-9364(2004)130:6(853).

Farha Shazmeen, S. *et al.* (2013) ‘Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis’, *IOSR Journal of Computer Engineering*, 10(6), pp. 2278–661. Available at: www.iosrjournals.org.

Grimes, S. (2007) *A Brief History of Text Analytics*. Available at: <http://www.b-eye-network.com/view/6311> (Accessed: 11 September 2017).

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques*. 3rd edn, San Francisco, CA, *itd: Morgan Kaufmann*. 3rd edn. Waltham: Morgan Kaufmann Publishers. doi: 10.1016/B978-0-12-381479-1.00001-0.

Hawkins, S. *et al.* (2002) ‘Outlier Detection Using Replicator Neural Networks’, in *International Conference on Data Warehousing and Knowledge Discovery*. Berlin, Heidelberg: Springer, pp. 170–180. doi: 10.1007/3-540-46145-0_17.

IHS Economics (2013) *2013 Global Construction Outlook*.

Jiang, J. (2012) ‘Information extraction from text’, in *Mining Text Data*. Springer Science+Business Media, LLC, pp. 11–41. doi: 10.1007/978-1-4614-3223-4_2.

Kensek, K. M. (2014) *Building Information Modeling*. 1st edn, *Handbook of Green Building Design and Construction*. 1st edn. Edited by R. E. Smith. New York: Routledge. doi: 10.1016/B978-0-12-385128-4.00005-6.

Khumalo, S. (2018) ‘SA Unemployment rate steady at 26.7%’, *Fin24*, 15 May. Available at: <https://www.fin24.com/Economy/Labour/sa-unemployment-rate-stays-steady-at-26-20180515>.

Kiprotich, C. J. K. (2014) *An Investigation on Building Information Modelling In Project Management: Challenges, Strategies and Prospects in the Gauteng Construction Industry, South Africa*. University of the Witwatersrand.

Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006) ‘Data preprocessing for supervised learning’, *International Journal of Computer Science*, 1(2), pp. 111–117. doi:

10.1080/02331931003692557.

Lee, J., Hsueh, S. and Tseng, H. (2008) ‘Utilizing data mining to discover knowledge in construction enterprise performance records’, *Journal of Civil Engineering and Management*, 14(2), pp. 79–84. doi: 10.3846/1392-3730.2008.14.2.

Lee, S. *et al.* (2011) ‘Data mining-based predictive model to determine project financial success using project definition parameters’, *28th International Symposium on Automation and Robotics in Construction, ISARC 2011*, pp. 473–478. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84863753135&partnerID=40&md5=0f5397513060366252a2792f36cecc00>.

Liaw, A. and Wiener, M. (2002) ‘Classification and Regression with Random Forest’, *R News*, 2(3), pp. 18–22. doi: 10.1177/154405910408300516.

Loper, E. and Bird, S. (2004) ‘NLTK: The Natural Language Toolkit’, in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, p. 31. doi: 10.3115/1118108.1118117.

Mcafee, A. and Brynjolfsson, E. (2012) ‘Big Data: The Management Revolution’, *Harvard Business Review*, October. Available at: <https://hbr.org/2012/10/big-data-the-management-revolution>.

Mukuka, M., Aigbavboa, C. and Thwala, W. (2015) ‘Effects of Construction Projects Schedule Overruns: A Case of the Gauteng Province, South Africa’, *Procedia Manufacturing*. Elsevier B.V., 3(Ahfe), pp. 1690–1695. doi: 10.1016/j.promfg.2015.07.989.

Nadeau, D. and Sekine, S. (2007) ‘A survey of named entity recognition and classification’, *Linguisticae Investigationes*, (30), p. 3–26. doi: 10.1075/li.30.1.03nad.

Nayak, R. and Qiu, T. (2005) ‘A data mining application: Analysis of problems occurring during a software project development process’, *International Journal of Software Engineering and Knowledge Engineering*, 15(04), pp. 647–663.

NBS (2014) ‘NBS National BIM Report 2014’, *RIBA Enterprise Ltd*, p. 36. doi: 10.1017/CBO9781107415324.004.

Niazi, G. A. and Painting, N. (2017) ‘Significant Factors Causing Cost Overruns in the Construction Industry in Afghanistan’, *Procedia Engineering*. The Author(s), 182, pp. 510–517. doi: 10.1016/j.proeng.2017.03.145.

Nicholas, J. M. and Steyn, H. (2012) *Project Management for Engineering, Business and Technology*. 4th edn. Milton Park, Abingdon: Routledge.

Pedregosa, F. *et al.* (2011) ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research*, 12, pp. 2825–2830. doi: 10.1007/s13398-014-0173-7.2.

Pospieszny, P. (2017) ‘Application of Data Mining Techniques in Project Management – an Overview’, *CEA Annals*, (43), pp. 199–220.

Redman, T. C. (1998) ‘The impact of poor Data quality on Typical a Enterprise’, *Communications of the ACM*, 41(2), pp. 79–82.

- Schoonwinkel, S., Fourier, N. and Conradie, P. (2016) 'A risk and cost management analysis for changes during the construction phase of a project', *Journal of the South African Institution of Civil Engineering*, 58(4), pp. 21–28. doi: 10.17159/2309-8775/2016/v58n4a3.
- Senouci, A., Ismail, A. and Eldin, N. (2016) 'Time Delay and Cost Overrun in Qatari Public Construction Projects', *Procedia Engineering*. The Author(s), 164(June), pp. 368–375. doi: 10.1016/j.proeng.2016.11.632.
- Smith, D. and Tardif, M. (2009) *Building Information Modelling: A Strategic Implementation Guide for Architects, Engineers, Constructors, and Real Estate Asset Managers*. John Wiley & Sons.
- Srivastava, S. (2014) 'Weka-01: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining', *International Journal of Computer Applications*, 88(10). doi: 10.5120/15389-3809.
- Sundheim, B. (1991) 'Overview of the Third message understanding evaluation and conference', *Proceedings of the 3rd conference on Message understanding*, 298, pp. 3–16. doi: 10.3115/1071958.1071960.
- Syvajarvi, A. and Stenvall, J. (2010) *Data Mining in Public and Private Sectors: Organizational and Government Applications: Organizational and Government Applications*. 1st edn. Edited by K. Klinger et al. Hershey: Information Science Reference. Available at: https://books.google.com/books?id=sI2fcLf_rV8C.
- Tang, W. *et al.* (2007) 'Risk management in the Chinese construction industry', *Journal of Construction Engineering and Management*, 133(12), pp. 944–956. doi: 10.1061/(ASCE)0733-9364(2007)133:12(944).
- Tharp, J. (2012) 'Project management and global sustainability', in *PMI® Global Congress - EMEA, Marsailles, France*. Newtown Square, PA: Project Management Institute.
- The Standish Group (2014) 'The Standish group: the chaos report', *Project Smart*, p. 16. doi: 10.1016/S0895-7061(01)01532-1.
- Tixier, A. *et al.* (2016) 'Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports', *Automation in Construction*, 62, pp. 45–56.
- Williams, T. P. and Gong, J. (2014) 'Predicting construction cost overruns using text mining, numerical data and ensemble classifiers', *Automation in Construction*. Elsevier B.V., 43, pp. 23–29. doi: 10.1016/j.autcon.2014.02.014.
- Witten, I. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd edn. San Francisco: Morgan Kaufmann Publishers. doi: 0120884070, 9780120884070.
- Zaki, M. J. and Meira, W. J. (2014) *Data Mining and Analysis: fundamental concepts and algorithms*. 1st edn, *Data Mining and Analysis*. 1st edn. New York: Cambridge University Press.

9 Appendix 1: Literature Review Summary:

The literature review in Chapter 2 focused on the importance of basing decisions on accurate data, on case studies of data mining in the construction sector, and on case studies of data mining to improve project management in the software development sector and the construction project environment. Information from all of these areas was used to synthesise a data mining process specifically for project management in the construction sector, thus fulfilling Objective 1 of Aim 1 (see Section 1.2) of the research.

Value of Data: The value of accurate data on which to base decisions was discussed. Data-driven decisions have resulted in increased efficiency, productivity and profitability for organisations embracing data mining and analysis. The negative impact of poor data affects all levels of an organisation and project equally. Equally, poor quality data will negatively affect the outcomes of data-driven decisions if care is not taken to solve data quality issues. Therefore, organisations that utilise data quality assurance techniques will experience tangible benefits, such as increased profitability and revenue generation, as well as intangible benefits, such as improved worker moral.

Case Studies: Data Mining in Construction: 5 case studies of data mining in the construction sector were examined. The case studies employed a variety of data mining techniques to fulfil specific goals determined at the outset of their investigation. The data mining processes used differed from application to application, but major themes were extracted from these case studies that were used in the synthesis of the data mining process.

The importance of setting a clear goal, of collecting suitable data (textual and numeric), of cleaning and preparing the dataset, of selecting and applying suitable data mining algorithms and validating the results of the application are common among the 5 case studies.

Case Studies: Data Mining in the Software Development Sector: 3 case studies of data mining in the software development sector were examined to compensate for the absence of case studies in the construction sector aimed specifically at improving project management. A relatively complete data mining process was presented by Case Study 6. This process represented the single largest influence in synthesising an applicable data mining process.

The other case studies supplied valuable information about how and when data mining can be used to improve project management. Reducing uncertainty and risk during the 4 project phases by providing accurate information is a key insight into improving project management.

Construction Project Environment: The construction project lifecycle was examined to determine if construction projects progress through the same phases as software development projects and if the accompanying uncertainty and risk levels were similar. It was found that construction projects have 5 phases in their lifecycle, instead of the 4 of software development projects, but that there are many similarities between them. This included the levels of uncertainty and risk associated with each phase.

The five sustainable construction project management criteria of Cost, Time, Quality, Environmental Impact and Social Impact were examined. Successful projects are able to balance these interlinked criteria. Each of these criteria have their own associated uncertainties and risks throughout the project. Therefore, project management can be improved by reducing the uncertainties or risks in any or all of them.

Suitability of Applying Data Mining to Project Management in the Construction Sector:

In terms of project phases and uncertainty levels, construction projects were shown to have sufficient similarities to software development projects to suggest that data mining to facilitate project management will be appropriate. Similarities were also shown between the two sectors in terms of the use of accurate information to reduce uncertainty or risks within a project phase.

Finally, data availability was discussed. The large amount of data typically stored in project databases along with the significant amount of documentation generated for each project provides enough data about the project throughout its entire lifecycle to enable data mining at each phase of the project.

Synthesis of a Data Mining Process: The information from the 5 case studies of data mining in construction was combined with information from the 3 cases studies of data mining in software development and with other knowledge about construction projects to synthesis a data mining process. The data mining process described was specifically tailored for application in the construction sector to assist project management and may be followed from the initial formation of an application goal through to validation of the results. These models are designed to produce output that may be used to reduce uncertainties in projects and to make wiser, data-driven decisions.

10 Appendix 2: Chapter 3 Summary:

Chapter 3 discussed the data mining process defined in Chapter 2 by drawing on information from the construction project environment discussed in Section 2.5 and three leading data mining textbooks. Two data mining examples, one real and one fictitious were examined to demonstrate the data mining process. The discussion of the data mining process fulfilled the second of the four goals set for the research.

Goal Definition: The importance of a clearly defined goal for data mining applications was discussed. The process for setting a goal when data mining to facilitate project management in the construction sector was also discussed. The main goal of the application should aim to reduce the uncertainty and risk exposure of a project by providing information during the project lifecycle. The phases of greatest uncertainty and risk i.e. the initiation and planning phases, are where the greatest contributions to project management are likely to be made.

Data Collection: Data can be collected for data mining from a variety of sources. The sources discussed were project databases, project documentation and building information models. Project databases are common in large engineering firms, government organisations, large contractors and private client bodies. These data bases are the easiest and most accessible source of data for data mining. Data can be gathered from project documentation by applying information extraction techniques. This data can then be used to supplement project databases. The typical documentation within a South African construction project was presented to show the wide variety and the spread of the documentation typically stored by the creator and the recipient. The possibilities of using BIM as a source of data in the future was also discussed.

Data Pre-processing: The data pre-processing step contains two sub-steps, namely goal re-examination and data pre-processing. During the goal re-examination sub-step, the investigator must determine if the goal of the application is feasible with the data provided. The pre-processing sub-step contains three distinct techniques that can be used to prepare the target dataset. These techniques can be employed simultaneously to produce a clean, normalised dataset that contains informative features and is representative of the data environment.

Data Mining and Modelling: Data mining algorithms typically function by determining the similarity between data points. Association pattern mining is an old form of data mining that is still used. It generates association rules which group reoccurring items based on their probability of occurring together. Similarity and distance functions serve as the basis of clustering, outlier analysis, classification, and regression models. Clustering and outlier analysis aim to group unlabelled data points into groups of similar data or to label them as outliers if they do not follow the internal data patterns. Classification aims to provide a class label to a data point after learning the internal data patterns from a labelled training dataset. Regression aims to provide a continuous value as the label for a data point, like regular regression, after training on a labelled dataset. Text mining is a unique form of classification that requires special data pre-processing steps and modified data mining algorithms.

Validation and Evaluation: Before a data mining model can be used in practice, it must first be tested to determine if the results are valid and accurate. The validation of clustering is a difficult task as a test sample usually does not exist. Therefore, clustering employs internal or synthetic external validation techniques to determine if the results are valid. Outlier analysis suffers from the same problem as clustering analysis in that a test sample rarely exists. Internal validation criteria are too biased to accurately validate results. Problems do exist with synthetic external validation techniques but they are utilised because they are more reliable than internal validation techniques. Classification models are the easiest to evaluate. Several performance measures exist to measure the overall and class specific accuracy of a classifier. These performance measures are used within one of several evaluation techniques that aim to provide reliable validation statistics.

11 Appendix 3: Data Mining Resources Summary:

Chapter 4 focused on the data mining resources that are available. Five possible resources were presented based on their completeness while not sacrificing mathematical rigor. The list of resources is not an exhaustive list. There are many other resources that aim to provide user-friendly interfaces with comprehensive data mining abilities.

The resources presented were selected based on how comprehensive they are and their ease of use. The resources are split between those that require the implementation to be done in a programming environment and those that are self-contained software packages. The resources that require programming knowledge are:

- **Machine learning in R (mlr):** This package for the R statistical programming language provides a user-friendly implementation framework for a large number of data mining and machine learning models. The package provides the required supporting methods, such as data preparation and algorithm evaluation techniques, to be a fully functional independent data mining resource.
- **Natural Language Toolkit (NLTK):** The NLTK module for the Python programming language is a comprehensive natural language processing module. The module provides a large toolkit for processing and examining text. The module provides some built-in classification capabilities but supports “wrapping” the processed data into a format that a sophisticated machine learning module, such as Scikit-learn, can utilise.
- **Scikit-learn:** This Python module is focused on providing a comprehensive machine learning environment that is well documented, with explanations and implementation guides for every method within the toolkit. The module contains a wide variety of methods that can be employed throughout the entire data mining process.

The self-contained software packages presented are:

- **Orange:** Orange is a free-to-use software package that provides a wide range of machine learning models and techniques that can be applied at every step of the data mining process. The environment utilises a visual programming interface that increases the user-friendliness and lowers the bar to entry for individuals with limited programming skills.
- **RapidMiner:** RapidMiner is a subscription-based software package that also utilises a visual programming interface. The software package provides over 1500 data processing, data visualisation, and modelling and validation methods to form a comprehensive toolkit for all possible data mining requirements. The package also provides the AutoMiner functionality that automates much of the data mining process.

The resources were evaluated for their ability to fulfil the requirements of the data mining examples given in Chapter 3. Four of the resources were able to perform nearly all the

required functions with the only drawback being they are unable to perform information extraction.

These resources, while not directly connected to data mining to aid project management in the construction sector, do provide the tools to enable construction sector personnel such as engineers and project managers to perform data mining on project information.

Table 11-1: Data Mining Resource Evaluation

		Requirements:		Resources:			
Step:	Method:	Scikit-learn:	NLTK:	MLR:	Orange:	RapidMiner:	
Application 1: Data mining to predict cost overruns.	Pre-processing	Normalising	x		x	x	x
		Data Imputation	x			x	x
		Wrapper Feature Selection	x		x	x	x
	Regression Models	SVM	x		x	x	x
		Decision Tree	x		x	x	x
		Multiple Linear Regression	x		x	x	x
		K-Nearest-Neighbours	x	x	x	x	x
		Neural Network	x		x	x	x
	Validation	K-Fold Cross Validation	x		x	x	x
		R	x		x	x	x
		MAE	x		x	x	x
		RMSE	x		x	x	x
	Application 2: Data mining to predict the impact of an on-site accident	Data Acquisition:					
		Information Extraction		x			
Pre-processing		Normalising	x		x	x	x
		Data Imputation	x			x	x
		Filter Feature Selection	x		x	x	x
Classification Models		SVM	x		x	x	x
		Naïve Bayes	x	x	x	x	x
		Decision tree	x	x	x	x	x
Validation		K-Fold Cross Validation	x	x	x	x	x
		Accuracy	x	x	x	x	x
	Recall	x		x	x	x	

12 Appendix 4: Data Mining Implementation

Table 12-1: Sample of Objective Transformation

ID	Objective	Objective binned
163	Upgr/Cap&Geo Imp Const Surf (Cap)	13
168	Reconstr/Rehab Constr Surf (Cap)	7
191	Reseal/Resurfacing (Cap)	9
195	Reconstr/Rehab Constr Surf (Cap)	7
197	Upgr/Cap&Geo Imp Const Surf (Cap)	13
200	Upgr/Cap&Geo Imp Const Surf (Cap)	13
203	Bridge Main Surf (Cap)	3
206	Reconstr/Rehab Constr Surf (Cap)	7
207	Reconstr/Rehab Constr Surf (Cap)	7
208	Upgr/Cap&Geo Imp Const Surf (Cap)	13
209	Reseal/Resurfacing (Cap)	9
212	New Roads Constr Surf (Cap)	5

Table 12-2: Tender Cost Estimate Standardisation

ID	Cost Estimate:	Cost Standardised:
2704	R 73 478 618,18	0,358150496
2831	R 47 591 836,75	0,231818051
2832	R 33 081 763,74	0,161006118
2833	R 27 614 402,76	0,134324351
2835	R 71 537 080,51	0,348675421
2837	R 21 021 200,97	0,102148265
2842	R 3 046 110,21	0,014426383
2844	R 2 584 319,69	0,012172757
2845	R 7 235 290,20	0,034870383
2846	R 3 532 753,00	0,016801293
2847	R 6 447 707,79	0,03102683
2848	R 2 254 747,87	0,010564384
2849	R 2 390 476,37	0,011226765
2850	R 10 950 001,12	0,052998883
2851	R 2 575 126,92	0,012127895
2852	R 3 883 112,57	0,018511115

Table 12-3: Employment Opportunities Binning:

ID	Employment Opportunities	Employment Opportunities Binned
163	180	3
197	195	3
200	211	4
203	40	0
207	143	2
245	105	2
262	213	4
295	288	5
306	36	0
327	109	2
334	157	3
335	123	2
336	19	0
340	35	0
352	97	1
353	161	3

Table 12-4: Target Dataset Sample

ID	District (Binned)	Objective (Binned)	Cost (Standardised)	Jobs Created (Binned)
163	1	13	0.788202745	3
197	2	13	0.670473726	3
200	1	13	0.172026677	4
203	3	3	0.030122941	0
207	4	7	0.350160809	2
245	1	7	0.278951735	2
295	2	11	0.357988627	5
306	4	13	0.056558279	0
334	1	12	0.163233136	3
335	5	7	0.074190098	2
336	4	3	0.049045213	0
340	2	7	0.079451466	0
352	6	1	0.07437963	1
353	5	3	0.18447622	3
368	2	6	0.016883495	0
381	1	7	0.339514909	4

```

#Importing Dependents
import numpy as np
import pandas as pd
from sklearn import model_selection, neighbors, svm, tree, naive_bayes, neural_network
#Setting up target dataset
df = pd.read_csv('Processed data.csv')
X = np.array(df.drop(['Jobs binned'], 1))
y = np.array(df['Jobs binned'])
#Initialising Classifiers
clf = neighbors.KNeighborsClassifier()
clf1 = naive_bayes.GaussianNB()
clf2 = naive_bayes.MultinomialNB()
clf3 = svm.SVC()
clf4 = tree.DecisionTreeClassifier()
clf5 = neural_network.MLPClassifier()
acc, acc1, acc2, acc3, acc4, acc5, n = 0, 0, 0, 0, 0, 0, 10

#Conducting 10 fold holdout validation process
for i in range(n):
    X_train, X_test, y_train, y_test = model_selection.train_test_split(
        X, y, test_size = 0.20, random_state = i)

    clf.fit(X_train, y_train)
    acc += clf.score(X_test, y_test)

    clf1.fit(X_train, y_train)
    acc1 += clf1.score(X_test, y_test)

    clf2.fit(X_train, y_train)
    acc2 += clf2.score(X_test, y_test)

    clf3.fit(X_train, y_train)
    acc3 += clf3.score(X_test, y_test)

    clf4.fit(X_train, y_train)
    acc4 += clf4.score(X_test, y_test)

    clf5.fit(X_train, y_train)
    acc5 += clf5.score(X_test, y_test)

#Printing Results
print("Neighbours: ", acc/n)
print("Naive Bayes: ", acc1/n)
print("Multinomial Bayes: ", acc2/n)
print("SVC: ", acc3/n)
print("Decision Tree: ", acc4/n)
print("Neural Network: ", acc5/n)

```

Figure 12-1: Implementation of Data Mining Application (Section 5.2)