Application of Data Mining techniques to identify significant patterns in the Grade 12 results of the Free State Department of Education

by
Aubrey Monde Madiba

Thesis presented in fulfilment of the requirements for the degree of Master of Philosophy (Information and Knowledge Management) in the Faculty of Arts and Social Sciences at Stellenbosch University

Supervisor: Ms Heidi van Niekerk

March 2017

DECLARATION

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

Signed Date: March 2017

Abstract

Application of Data Mining techniques to identify significant patterns in the Grade 12 results of the Free State Department of Education

Aubrey Monde Madiba

Department of Information Science

University of Stellenbosch

Thesis: Master of Philosophy (Information and Knowledge Management)

March 2017

The Free State Department of Education (FSDoE) has a mandate to ensure that examinations and assessment processes are conducted according to the set out legislations and that they produce expected results. It has become common for Grade 12 results to be challenged by interested parties within and outside the government on their credibility. It is, therefore, the responsibility of the Free State Department of Education to ensure that the input data which represent raw marks obtained by the learners give a true reflection of what individual learners have achieved during a particular assessment period.

This study seeks to explore the role that data mining (DM) can play in establishing credibility of the Grade 12 data in the FSDoE. The study makes use of open-source data mining software called WEKA. The software is applied on the 2010-2013 Grade 12 data results in the Free State. For this study, two algorithms, *j48*, and *simpleKMeans* algorithms, have been selected for classification and clustering respectively. In line with the universally accepted Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, the selected data has been modified and saved in WEKA software-compliant csv format.

The prepared data represent four selected subjects which are English Home Language (EHL), English First Additional Language (EFAL), Mathematics and Mathematical Literacy. Four Different models were iteratively generated and analysed and valuable insights were drawn from them to highlight how their possible influence on future decision making in the FSDoE. The analysis focuses on performance of learners within the performance categories (levels 1 to 7) and compares them Free State's Grade 12s average performance during the selected 2010 to 2013 period. The English Languages (EHL and EFAL) models and the Mathematics (Mathematics and Mathematical Literacy)

models are analysed and interpreted according to the identified patterns as observed over the four year period (2010-2013).

In addition, the study makes sense of the models generated from WEKA by interpreting them using theories from Bloom's Mastery Learning and Argyris' Learning Organisations. Furthermore, the study delves into the 2011 census data and make sense of the results obtained from the application of WEKA in the selected 2010-2013- Grade 12 results in the FSDoE.

The study concludes by giving recommendations which the Free State Department of Education may use as they plan not only for future Grade 12 results but across all grades. It is through the application of DM tools that credibility, as seen with Grade 12 data in the FSDoE, can be established through sense making which can assist during decision making.

Opsomming

Aanwending van Data-ontginnings tegnieke om betekenisvolle patrone te identifiseer in die Graad 12-resultate van die Vrystaatse Onderwysdepartement

Aubrey Monde Madiba

Departement Inligtingwetenskap

Stellenbosch Universiteit

Proefskrif: Magister in Wysbegeerte (Inligting en Kennisbestuur)

Maart 2017

Die Vrystaatse Onderwysdepartement (VOD) het 'n mandaat om te verseker dat eksamens en assessering prosesse volgens die uiteengesette wetgewing uitgevoer word en dat hulle verwagte resultate produseer. Dit is deesdae algemeen dat Graad 12-uitslae uitgedaag word deur belanghebbende partye binne en buite die regering ten opsigte van geloofwaardigheid. Dit is dus die verantwoordelikheid van die VOD om te verseker dat die rou punte wat deur leerders behaal word 'n ware weerspieëling van individue se prestasie tydens 'n assesserings periode is.

Hierdie studie beoog om die rol van data-ontginning in die bepaling van die geloofwaardigheid van Graad 12-data in die Vrystaat te ondersoek. Die studie maak gebruik van WEKA, ’n publieke data-ontginningsagteware pakket. Die sagteware word toegepas op 2010-2013 se Graad 12 resultate in die Vrystaat. Vir hierdie studie sal twee algoritmes, j48, en simpleKMeans, onderskeidelik vir klassifikasie en groepering gebruik word. Die data is bygewerk en in csv formaat volgens CRISP-DM metodologie gestoor.

Die bygewerkte data verteenwoordig vier geselekteerde vakke wat Engels Huistaal (EHT), Engels Eerste Addisionele Taal (EEAT), Wiskunde en Wiskundige Geletterdheid insluit. Vier modelle is iteratief gegenereer en ontleed wat interessante insigte met ’n impak op toekomstige besluitneming van die VOD gelewer het. Die analise fokus op die prestasie van leerders binne die prestasie kategorieë (vlakke 1-7) en vergelyk dit met Vrystaat se gemiddelde prestasie tydens die gekose 2010-2013 tydperk. Die Engelse taal modelle (EHT en EEAT) sowel as die Wiskunde modelle (Wiskunde en Wiskundige Geletterdheid) is volgens die geïdentifiseerde patrone, soos waargeneem oor die tydperk van vier jaar (2010-2013), ontleed en vertolk.

Daarmee saam het die studie sin gemaak van die Weka gegenereerde modelle en met behulp van Bloom se Bemeesterings Leerteorie en Lerende Organisasies soos opgevat deur Argyris

geïnterpreteer. Verder maak die studie gebruik van die 2011-sensus data om meer insigte oor die gegenereerde modelle wat die Graad 12-resultate van die VOD te bekom.

Ten slotte maak die studie aanbevelings vir die VOD wat hulle kan gebruik vir die beplanning van nie net die toekomstige Graad 12 eksamens nie, maar in alle grade. Met sinvolle toepassing van data-ontginningsagteware tydens besluitneming kan geloofwaardigheid, soos gesien met Graad 12 data van die VOD, vasgestel word.

Acknowledgements

A word of thanks goes to God, Almighty, who has been with me throughout this period. During the time of doubt and wanting to give up, He raised me up so that I could stand on mountains, walk on stormy seas and I got stronger because I knew I was safe on His shoulders[1].

I am greatly indebted to my supervisor Ms. Heidi van Niekerk whose insights, knowledge, patience, guidance, and humanity made me believe in this project. Your words of encouragement and "straight talk breaks no friendship" approach are the reasons why I was able to see the finish line. I will forever be indebted to you. To the entire MIKM lecturers, thank you for patiently taking us through the whole Information and Knowledge Management landscape and made us believe we had the ability to be the change-makers in the knowledge economy.

Furthermore I would like to thank Mr Phosa from FSDoE who was there for me when I needed the Grade 12 data and from day one he understood the reasons why such data had to be released.

To my family and friends I hope you now understand why this had to be done. I set out on a mission and it was up to me to either fulfil it or betray it. I chose the former and it bore fruit. To my late parents I dedicate this piece of work to you for without the solid foundation and unconditional love, I would not have managed to traverse this challenging terrain.

[1] (Groban 2001)

Table of Contents

List of Figures

List of Tables

List of Graphs

Chapter 1 *Introduction*

1.1 Background of the study

The Free State Department of Education conducts an examination and assessment for Grade 12 learners through its examinations and assessment directorate, annually. The directorate's functions are guided by, among other regulations, the General and Further Education and Training Quality Assurance Act No. 58 of 2001 which defines assessment as:

> The process of identifying, gathering and interpreting information about a learner's achievement in order to
>
> (a) assist the learner's development and improve the process of learning and
>
> (b) evaluate and certify competence in order to ensure qualification credibility[2].

In line with this study, this Act calls for the collection of data about the learners' progress which must be stored in a reliable source with a view to using it to improve teaching and learning in schools. Tied to this Act is the Bill of Rights that clearly calls for the collection of data to be conducted within the ambit of law. According to the Bill of Rights as enshrined in the Constitution of the Republic of South Africa, Act no. 108 of 1996, 'everyone has the right to a basic education… which the state, through reasonable measures, must make progressively available and accessible (South Africa. Constitution of the Republic of South Africa No 108 of 1996 1996)[3].

As stated in the Bill of Rights, this means that the state has the responsibility to provide a well-structured assessment and examinations infrastructure that will ensure that, as learners progress through the grades, they will know with confidence that the data representing their marks is a true reflection of their abilities. Their progress data reports, which are collected over many years, should be stored in a reliable database which will indicate their levels of achievement and performance over time. In addition, this indicates the importance of having a detailed knowledge base that would reflect the abilities of the learners. This knowledge base should also, from time to time, grant access to interested parties, to analyse and interpret its data in order to assist the state to make informed decisions about its citizens.

The end of the Grade 12 year is a milestone in the lives of many Grade 12 learners as it marks the end of a gruelling 12-year school career. On the other hand, it is the beginning of a new life in a

[2] (South Africa. General and Further Education and Training Quality Assurance Act No 58 of 2001. 2001)

[3] (South Africa. Constitution of the Republic of South Africa No 108 of 1996 1996)

tertiary institution or working environment. During the release of results, more emphasis is placed on the overall percentage pass rate by schools or provinces, and little, if any, on the individual learning areas. The excitement soon subsides and reality sets in when individual learners have to present their individual results to their respective tertiary institutions for admission. The main frustration occurs when they are informed that they have not met the often strict entrance requirements. The individual marks are usually weighed against the faculty's requirements. The study, therefore, tries to determine whether there are any patterns that can be identified from the raw data that is generated every year during the Grade 12 examinations in the Free State. The application of data mining (DM) tools and techniques on the data obtained from selected subjects in the database determines whether the tool is applicable to an examinations and assessment environment. This study makes use of the FSDoE's database which is administrated by both Umalusi and Education Management Information System (EMIS).

According to the General and Further Education and Training Quality Assurance Act No. 58 of 2001, Umalusi has been given the responsibility of upholding the quality of all the areas that affect the examinations and assessment process in the education system. More specifically, Paragraph 16(2) (e) of this Act states that Umalusi has the main task of issuing certificates for qualifications at the exit points in the General and Further Education and Training bands. Umalusi is further mandated to ensure that these certificates are credible both nationally and internationally[4]. For the certificates to obtain a stamp of approval, the processes of data collection and interpretation in all areas of learning are of crucial importance and, as a quality assurer, Umalusi will have to make its presence felt.

As a publicly-funded institution, Umalusi agrees that the value that the public places on the external examinations in any country is solely dependent on well-set-out education standards which are not only simple, reliable and easy to understand, but which are also attainable[5]. To achieve such well-thought-out objectives is always a challenge and, in many instances, Umalusi has been found wanting, trying to find answers regarding the controversial processes that it applies in dealing with examinations and assessment data. There are many instances in which Umalusi's quality assurance operations have been put to the test.

In one of their Newsletters called *Makoya*, Umalusi admitted that standardisation, as one of their core functions, is still an elusive and less-understood concept to many[6]. They argue that in 2006, for

[4] (Umalusi. Directives for certification. National Senior Certificate (schools) 2008)

[5] (Umalusi. Quality Assurance of Assessment: policies, directives, guidelines and requirements 2006)

[6] (Umalusi. The standardisation of the final examinations 2007)

example, the raw marks of different subjects went through the standardisation process which is handled by a committee comprised of prominent people who are knowledgeable about standardisation. Despite Umalusi's explanation, it seems that there is a veil of secrecy surrounding the tools that they are using to produce credible and acceptable Grade 12 results on the Grade 12 level[7].

According to the Parliamentary Monitoring Group (PMG), parliamentarians raised their concerns with regard to the way in which the quality assurance body conducts its business[8]. With specific reference to the 2010 Grade 12 results that the parliamentarians questioned, 'Umalusi paradoxically maintained that standardisation was both confidential and not a secret'[9]. They further stated that standardisation was an internationally-observed process aimed at ensuring that learners were neither advantaged nor disadvantaged by factors other than knowledge of the subject and aptitude. The monitoring group was cautions of the fact that the standardisation process needed to be handled with care as failing to do so could result in it being wrongly interpreted[10].

Andrew Trench adds that what makes Grade 12 results data questionable is the fact that the stakeholders cannot prove convincingly and in detail the quality of these results[11]. During his investigation of the 2010 Grade 12 results, Trench observes that on their own, numbers mean nothing unless massaged well enough to give out what he calls a simple – and even painful – truth (Trench, Andrew 2012)[12].

It is as a result of the above concerns and investigations that this study, which involves the application of DM tools and techniques, was undertaken. The application of DM tools and techniques in a study such as this one may compel Umalusi to play open cards as transparency is an important tool for gaining public confidence with regard to controversial issues involving Grade 12 examinations and assessment data. Unless it is collected on time and makes use of the correct tools, the examinations and assessment data will forever be difficult to understand. It is on controversial matters such as these that this study intends, through the application of DM tools and techniques processes, to determine

[7] (Umalusi. The standardisation of the final examinations 2007)

[8] (Parliamentary Monitoring Group. National Senior Certificate Examinations 2010: briefing by the department and Umalusi 2010)

[9] (Umalusi. The standardisation of the final examinations 2007)

[10] (Parliamentary Monitoring Group. National Senior Certificate Examinations 2010: briefing by the department and Umalusi 2010)

[11] (Trench, Andrew 2012)

[12] (Trench, Andrew 2012)

whether the examinations and assessment data can be regarded as credible over a period of time. Given the challenges with regard to understanding data, alternatives need to be explored in order to restore the public's confidence regarding the Grade 12 results.

Due to the ever-increasing and overwhelming data, the need arose for data mining tools and techniques that are not only able to help us analyse data but which also lead to the production of valuable information needed by decision makers[13]. The development of new technologies, such as Knowledge Discovery in Databases (KDD), assists the human mind in discovering valuable information through data analysis. This development originated in response to the older statistical techniques that have proved to be not only out-dated but also unable to handle the massive data produced by the new technologies[14].

Decision makers in many educational institutions are facing the mammoth task of finding the modern tools and techniques to help simplify the complexities associated with the massive amount of data with which they are confronted. They are forever looking for more efficient and effective data mining technologies to help them make better decisions and, in the process, develop new strategies for the future. By acquiring such technologies, they would be able to extract knowledge from both the historical and operational data found in their departments' databases. It is through DM tools and techniques that departments will manage to explore and uncover massive information that is inaccessible to the naked eye[15].

The South African education system operates in an environment where technologies form the back-bone of our daily operations. At both national and provincial levels, the education sector has a number of information systems which are either computerised or non-computerised. The presence of these systems has led to the creation of platforms which allow various departments to execute the various business activities with ease. These include, among others, the function of admitting learners into schools, registering learner attendance and achievement, closing and opening institutions, appraising educators, charging fees, communicating with parents, and so on. The Education Management Information System (EMIS), for example, and as seen later in this study, is the main provider of such raw data which is used during the process of data mining (South Africa. Department of Education 2005)[16].

[13] (Guruler, Istanbullu and Karahasan 2010)

[14] (Guruler, Istanbullu and Karahasan 2010)

[15] (M. Beikzadeh 2008)

[16] (South Africa. Department of Education 2005)

Trucano argues that the role of the Education Management Information System has always been the provision of information pertaining to education inputs such as the number of schools in a location, enrolment levels, and the number of teachers looking after pupils[17]. With such well-defined roles played by EMIS, including the handling of the examinations and assessment data, the absence of proper DM tools and techniques would still be regarded as a void that needs to be filled in our education system. As seen later, this study involves the application of data mining on data in which EMIS plays an important role, particularly with regard to its collection and storage. The application of unique DM tools and techniques, the focus of which is on educational data, serves an important function in interpreting the data.

Educational Data Mining (EDM), as an emerging field of study, comprises of a number of computational and research methods which assist researchers in obtaining more information on various issues related to the education sector. A number of such activities include the way in which students learn, and the environment in which learning takes place. EDM does not confine itself to learning about individual students, but also focuses on assessing their performance with the aim of improving the learning process after identifying barriers during the evaluation process[18].

Educational Data Mining can be described as both a learning science, as well as a rich application area for data mining. Due to the ever-increasing data related to educational matters, EDM has been identified as an enabler of data-driven decision making which, in turn, leads to an improvement in educational practice, and the provision and use of educational resources[19].

Beikzadeh adds that by extracting raw data to discover new knowledge, the Department of Education will be able to make informed decisions which would be supported by developed models which would have been uncovered during the application of DM tools and techniques. Having access to superior technologies that help to improve understanding with regard to the education data leads to the creation of reliable and easy-to-understand policies and procedures for the entire education sector[20].

Even though the presence of data in the education sector is often regarded as an ‘indicator of reality, or a measure of truth’, it has been observed that data in its raw state is always messy and tends to obscure the real facts[21]. Metaphorically speaking, data is sometimes likened to a light that triggers

[17] (Trucano 2006)

[18] (Bousbia and Belamri 2014)

[19] (Calders and Pechenizyky 2011)

[20] (M. Beikzadeh 2008)

[21] (Piety 2013)

action which then leads to illumination of all the dark areas[22]. Inconsistent data in the education sector cannot be discarded as a result of difficulties related to its interpretation. However, with the help of new technologies, corrections may be made to uncover valuable information which might have been hidden therein[23]. This study delves into this so-called messy data in the examinations and assessment database to uncover valuable information which may help educationists to gain a better understanding of both teaching and learning in the education sector.

Although the application of DM tools and techniques is the main reason behind undertaking this study, the role that Bloom's Mastery Learning model plays in understanding the outcomes of a process of assessment cannot be overlooked. A number of studies have shown that the model plays an influential role on various levels where teaching and learning take place, including the FSDoE under which Grade 12 learners fall[24]. Whatever valuable information is uncovered from the education data would need to be interpreted, using education-related models and theories.

This study on the FSDoE seeks to demonstrate the role of DM tools and techniques in improving our understanding of data generated from examinations and assessment processes by offering various models which will help the decision makers in the Department of Education to make informed decisions. The important thing about DM is its application of universally-accepted methodologies that are guided by the pre-set aims and objectives in each and every project.

This study, which focuses on the FSDoE's examinations and assessment database, follows an internationally-renowned CRISP-DM methodology which is discussed in detail later in this study.

By providing a detailed background on DM tools and techniques, this study aims to highlight its role in the whole 'knowledge discovery' process. This study, as indicated in the research question, intends to discover patterns from the FSDoE's examinations and assessment database. In order for that to happen, it is crucial that all the elements that contribute to the realisation of such an objective are highlighted. As a new field of study, DM tools and techniques need to be explored by highlighting their historical background. This is followed by a detailed account of the concept of DM tools and techniques as a guideline for the realisation of the objectives of this study. It is through understanding such concepts in their entirety that the later implementation of the DM tools and techniques in the FSDoE's examinations and assessment database is executed with ease and confidence.

[22] (Piety 2013)

[23] (Piety 2013)

[24] (Brown 2012)

1.2 Problem Statement

1.2.1 Research Question

Based on the controversies surrounding the Grade 12 results as highlighted above, the primary research question is:

How can the application of data mining tools and techniques assist in establishing credible Grade 12 results?

Subsidiary questions are:

How widespread is the use of Data Mining in the education sector?

What legislation guides the implementation of Data Mining tools and techniques in examinations' data?

Which influential institutions can help to sustain reliable Grade 12 results through the application of Data Mining tools and techniques?

Why is it important to select a good discovery tool when dealing with data?

What informs the selection of suitable Data Mining tools and techniques during the application of Knowledge Discovery in Databases?

What is the relationship between the reasons for applying Data Mining and the models generated thereafter?

How will the results of a Data Mining application contribute to the understanding of the state of education in a country?

1.2.2 Research Objectives

In order for South Africa's Grade 12 results to be credible, more attention needs to be paid to the Data Mining tools used to extract valuable information from the education database.

The objectives of this study are:

- to demonstrate the role played by Data Mining tools and techniques in establishing the credibility of the Grade 12 data which is stored in the education database;
- to highlight the influence of the models generated during the Data Mining activity on decision making processes in the Department of Education;
- to explain how the results obtained during data mining can be interpreted, using Bloom's Mastery Learning model;

- to highlight the role played by both single- and double-loop learning in influencing the outcomes of the Grade 12 data; and
- to establish whether there is a relationship between the results of the 2011 census and those obtained during the application of data mining tools and techniques to the Grade 12 results.

1.2.3 Research Approach

This study employs a comparative data analysis, using secondary data[25]. Hofstee argues that 'there is a huge amount of data available, scattered all over the world…as long as it is reliable, it has the potential'[26]. In this study, secondary data is from the FSDoE examinations and assessment database. Courtesy of the Education Management Information System (EMIS), the 2010 to 2013 Grade 12 data used in this study was copied and shared freely. In order for the above-mentioned data to be analysed and make sense out of it, reliable data mining software is needed.

This study uses the Waikato Environment for Knowledge Analysis (WEKA), a machine learning toolkit developed at the University of Waikato in Hamilton, New Zealand. The software provides machine learning, statistics and other data mining solutions for various data mining tasks such as classification, cluster detection, association rule discovery and attribute selection[27].

WEKA, which is written in Java and released under GPL (General Public Licence), was released in 1992 as a project funded by the government of New Zealand. It contains a number of popular machine learning methods which play an important role in statistical learning, but which are not typically found in statistical software packages. What makes WEKA important is not only its ability to provide a convenient and efficient platform but also its ability to provide data miners with software that allows them to create and compare results from different modelling algorithms[28].

It has been proven that no single machine learning method caters for all learning problems. This means that it is impossible to single out a machine learning method that has the ability to tackle various learning problems at once. The unique nature of datasets compels data miners to search for and select specific algorithms which assist them in solving the challenges they face. WEKA, a well-known state-of-the-art machine learning workbench, contains a number of algorithms. What makes WEKA stand out above the rest is its flexibility and ability to accommodate a variety of DM

[25] (Hofstee 2006)

[26] (Hofstee 2006)

[27] (Hofstee 2006)

[28] (Hofstee 2006)

experiments, from pre-processing to the evaluation of the results. WEKA's workbench provides a platform for solving a number of DM problems, using regression, classification, association rule mining and attribute selection[29].

CRISP-DM, which stands for Cross Industry Standard Process for Data Mining, is an initiative by a consortium of software vendors and industry, for the use of data mining technology to standardise the data mining process[30]. The methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. Smartvision Europe further adds that even though they are not its inventors, they are its evangelists because of its practicality, flexibility and usefulness when using analytics to solve thorny business issues. It is a golden thread that runs through almost every client engagement[31].

In the early years of data mining, many data miners used their own approaches and procedures to perform data mining. These approaches and procedures are heavily influenced by the nature of the input data and software tools. Quite often, trial and error was adopted in order to find the best solution after repeated attempts. By the mid-1990s, there was a strong desire in the data mining community for a methodology that is independent of industry, tool and application[32].

CRISP-DM was proposed by major data mining software vendors and practitioners who wanted an industrial standards process for data mining. The methodology proposes a thorough and rigorous methodology for undertaking data mining projects. It outlines the activities of data mining in six phases, consisting of a number of generic tasks[33].

Although WEKA has a variety of algorithms, for this study, only classification and clustering have been selected in order to assist in finding out whether there are patterns to be identified in the selected data. The two algorithms are popular and widely used in many studies similar to this one. They can be applied to both large and small datasets. J48 algorithm is used during classification, whereas SimpleKMeans is used during clustering.

For this study, the data from selected schools is identified, prepared and used during the application of DM tools and techniques. The subjects selected for this study include English Home Language (EHL) and csv, Mathematics and Mathematical Literacy. To ensure the credibility and reliability of

[29] (Rokach and Maimon 2005)

[30] (Refaat, Data Preparation for Data Mining using SAS. 2007)

[31] (Smart Vision-Europe: Predictive Analytics for Smarter Business 2015)

[32] (H. Du 2010)

[33] (H. Du 2010)

the results, the selected data is segmented into different data sets in order to perform a comparative analysis within a specified year, and across the selected years. The data to be used during classification is divided into training and test options. For clustering, the same data is divided into five clusters.

This study also employs the well-known model in education called Bloom's Mastery Learning which plays a critical role in the teaching and learning process[34]. The steps that Bloom's Mastery Learning theory uses are:

- Initial instruction
- Assessment
- Feedback and
- Corrective instruction[35].

Bloom's Mastery Learning, as a concept, was first introduced into many American schools in the 1920s. What affected its success and possible sustainability was the absence of suitable technology at the time. It was only when Bloom re-introduced it in the 1960s that the theory achieved widespread recognition. As a world-renowned theoretician and promulgator of Mastery Learning, Bloom predicted that in the classes where Mastery Learning is taught, as many as 95% of the students would achieve at the level that had previously been dominated by only 5%[36].

In addition, it has always been considered as a norm in many schools for teachers to expect a third of their students to pass, and another third to fail[37]. Such expectations of fixing the academic goals are not only wasteful and destructive, but also reduce the motivation for both teachers and learners to teach and learn, respectively[38]. Such a system further denies young people access to a variety of opportunities which are available for post-school learning. It should be noted that as many as 90% of students have the ability to master what is taught, and teachers have to look for various strategies that enable students to do so. Mastery Learning seeks to highlight those individual differences in learners which affect teaching and learning[39]. Such unfounded beliefs, that teachers currently hold, usually result in societies that do not aim for higher levels of success.

---

[34] (Davis and Sorrell 1995)

[35] (Davis and Sorrell 1995)

[36] (Davis and Sorrell 1995)

[37] (Bloom 1968)

[38] (Bloom 1968)

[39] (Bloom 1968)

The assumption that only a few members of society can be successful and serve the rest can be traced back to the beliefs held by teachers and examining bodies that a selected number of students qualify to be labelled as talented. More time is spent on the prediction and selection of talent, while little is spent on the development of such talent. Modern societies should come up with ways to include more students in the pool of success by devising strategies that will allow for effective learning through the provision of essential subject matter. This will be possible when the attitudes of teachers, students and administrators, as well as the strategies involved in teaching and evaluation, change[40]. Mastery Learning procedures allow teachers from all corners of the world to aim for success.

Mastery Learning is an essential instructional technique for teaching and learning which involves breaking down the subject matter into manageable units or lessons in which students are given time to learn, and are later tested. If they under-achieve, they are given additional teaching time until they achieve a mastery grade on the re-test. Mastery Learning emphasises the notion that achievement levels should be similar for all students, with the only difference being the time it takes to attain specified Mastery levels. The more equal time is given, the more there will be inequality of achievement[41]. The provided steps are a guarantee that the model can produce the expected results.

To be specific, Mastery Learning, allows teachers to break down the subject matter into manageable units which are taught according to the set objectives[42]. For students to master a unit, and before they can move on to the next one, they are expected to obtain 80% during exams. Those who fail to achieve that mark are afforded additional time for remediation[43]. Students continue the cycle of studying and being tested until mastery is attained. Mastery learning ensures that students who perform at a minimum level obtain a higher level of achievement than by means of traditional methods of instruction[44]. It is important that the educational environment is restructured and not only focus on achievement levels at specific groups of learners but look at an individual needs holistically including the time it takes for each learner to master specific concepts[45].

Furthermore, it should be noted that the main goal of Mastery Learning is for all students to achieve at higher levels, and that this is supported by students demonstrating a positive attitude and motivation

[40] (Arlin 1984)

[41] (Arlin 1984)

[42] (Davis and Sorrell 1995)

[43] (Davis and Sorrell 1995)

[44] (Davis and Sorrell 1995)

[45] (Davis and Sorrell 1995)

to learn. Its founder, Bloom believes that Mastery Learning improves students' attitudes and promotes an interest in effective learning[46]. The benefit of Mastery Learning lies in a solid foundation which makes it easier for learners to achieve higher levels later in their schooling. It has the ability to increase achievement across all subjects, and Mathematics as a subject that is feared by most, has a great potential for achievement due to its sequential and ordered nature[47].

Tied to Bloom's Mastery Learning is the concept of Learning Organisations as seen in Argyris' single- and double-loop learning. In order for schools to become centres of success, they need to address the concept of learning which seems to be elusive and difficult to understand. Learning is least understood because it is normally associated with 'problem solving' which involves the identification and correction of errors which exist in the outside world. The correct way of addressing the issue of learning is by reflecting critically on the way in which schools behave, identifying the causes of problems, and embarking on an attitude-changing journey[48].Distinguishing between the single and double-loop forms of learning will go a long way towards helping schools understand what learning is all about.

Argyris coined the two terms 'single-loop' and 'double-loop' learning in order to help institutions gain a better understanding of what learning is and what takes place during learning[49]. He explained this concept by using the analogy of a thermostat: it automatically turns up the heat whenever the room temperature drops to below 68 degrees. This he referred to as single-loop learning. He further states that if a thermostat could ask, 'Why am I at 68 degrees?' and then look for other alternatives which may be economical to heat the room that would be called double-loop learning[50].

The above analogy identifies single-loop learning as the process of making corrections whenever anomalies occur, and paying little attention to the values and prevailing factors which caused the disturbance. On the other hand, double-loop learning goes deeper and uncovers the root causes of the mismatch, and the skills needed to embark on such a journey are usually greater and more complicated than in single-loop learning[51]. During the application of single-loop learning, simple changes are effected, whereas double-loop learning involves a process of reframing which entails a different and

[46] (Davis and Sorrell 1995)

[47] (Davis and Sorrell 1995)

[48] (Argyris, Harvard Business Review: teaching smart people how to learn 1991)

[49] (Argyris, Harvard Business Review: teaching smart people how to learn 1991)

[50] (Argyris, Harvard Business Review: teaching smart people how to learn 1991)

[51] (Argyris, A life full of learning. 2003)

comprehensive way of doing things[52]. In addition, double-loop learning not only leads to changes in the existing operational activities, but also manifests itself as a tool of transformation that leads to the creation of new policies[53].

In addition, double-loop learning is not about whether we are 'doing things right' but whether we are 'doing the right things right'. In a teaching environment, this means moving away from using only the lecture method (single-loop) and employing a variety of teaching methods (double-loop). Such a paradigm shift with regard to the way in which teaching takes place not only produces good results but also leads to the converting institutions into Learning Organisations[54]. It is also common for educators with a limited cognitive frame to shift the blame onto students instead of inwardly looking at their attitudes, beliefs and behaviours. They tend to attribute their shortcomings to forces outside their spheres of influence and, in the process, block learning[55]. In contrast, if they were to apply double-loop learning, such teachers would engage in introspection which would involve changing their attitudes, values, beliefs and practice[56].

In order for schools to succeed, they need to focus on double-loop learning which allows them to fuse the current ways of doing things with new knowledge. Such an exercise leads to the creation of a new culture of doing things. A variety of strategies is often employed in the sharing of ideas, as well as in the expansion of the knowledge pool and the memory of an institution[57].

Self-evaluation tools include some modifications with regard to the way in which people do and apply things. This is done by gathering evidence, searching for the truth, differentiating between subjective and objective elements, and categorising assessment and evaluation into summative and formative to support what they are doing. Through this type of learning, schools will be able to change their image and become 'smart schools[58]'.

In addition, this study makes use of the 2011 Census in order to shed more light on the outcomes obtained from the interpretation of the results, using Bloom's Mastery Learning model. One important thing about a Census is that it is a solid source of demographic information at all levels of geography,

[52] (Georges 1999)

[53] (Georges 1999)

[54] (Mantz 2000)

[55] (Bensimon 2005)

[56] (Bensimon 2005)

[57] (Scribner, et al. 1999)

[58] (Pedder and MacBeath 2008)

on any given place and time. The results from the 2011 National Census provide interesting insights which have a direct influence on the data produced by the education sector on an annual basis. Since the dawn of democracy in South Africa, three censuses have been conducted (1996, 2001 and 2011). Censuses are an important tool for collecting data on issues pertaining to population, education and housing, which are crucially important during the creation of a national plan for socio-economic development, policy interventions which lead to their implementation, and evaluation at a later stage. The latest Census which was conducted in 2011, contains a number of important attributes which were first measured, and which led to the creation of a number of important indicators[59].

The focus is on the following:

- Population size: the focus is on the size of the population in the Free State in comparison with the rest of the country;
- Age-sex distribution: this concerns itself with the proportion of both men and women in the province and their respective ages;
- Race distribution: the composition of different racial groups are also highlighted in this study;
- Migration patterns: the study also examines how migration patterns both in and out of the Free State province are taking place;
- Schooling: this part examines the differences between the public and public schooling in the Free State province;
- Annual household income: the amount of money that individual households generate is also highlighted;
- Housing: the results on the types of housing structures found in the Free State are indicated in the study; and
- Provision of services: the nature of service delivery in the Free State is also highlighted and compared with that of the rest of the country[60].

The inclusion of the Mastery Learning model and the results of the 2011 Census assist in understanding the context in which teaching and learning take place in the schools falling under the Free State Department of Education.

[59] (Statistics South Africa. 2012)

[60] (Statistics South Africa. 2012)

1.3 Limitations

Hostee acknowledges that 'all methods have limitations. Your method's limitations are what separate doing your study according to your method from perfection. Perfection is seldom, if ever, attainable[61]'. In this study, a number of limiting factors have a direct influence on the way in which it is going to be conducted. One of these involves time.

Due to a limited timeframe in carrying out the research, the experimentation could not be applied on a wider scale. It has been narrowed down to the Grade 12 schools in one out of five possible districts in the Free State province, i.e. Motheo. To achieve a data comparison in this study, the same schools in the Motheo region are studied through the application of limited DM tools and techniques on the 2010 data which are to be compared with those obtained in 2011, 2012 and 2013 Grade 12 examination results. Financial constraints also make it impossible to use various data mining software available on the market as a way of comparing and testing the universality of the experiment's results.

[61] (Hofstee 2006)

Chapter 2 *Literature Review*

2.1 Definition of terms

2.1.1 Assessment

The National Protocol for Assessment defines assessment as 'a process of collecting, analyzing and interpreting information to assist teachers, parents and other stakeholders in making decisions about the progress of learners[62]'.

Assessment also refers to a judgement which can be justified according to specific weighted set goals, yielding either comparative or numerical ratings[63].

For the purposes of this study, the application of DM tools and techniques to Grade 12 data for the Free State plays an important role in highlighting how the process of assessment is carried out.

2.1.2 Credibility

Credibility can generally be defined as an act of believing at some point in time and is composed of trustworthiness and expertise. The two components are a reflection of a pattern that can be traced back over a period of time[64]. The Webster's New Collegiate Dictionary defines credibility as the act of offering 'reasonable grounds to be believed[65]'. Credibility allows for the justification of a developed model as a valid tool that can be used for research and making informed decisions[66].

For this study, models were developed by means of which the credibility of the FSDoE's Grade 12 results could be confirmed.

2.1.3 Data

Zimmermann defines data as:

[62] (South Africa. Department of Basic Education 2012)

[63] (Taras 2005)

[64] (Erdem and Swait 2004)

[65] (Meyer 1988)

[66] (Rykiel 1996)

> structured symbols, numbers, letters, or even words without any specific interpretation, which can be manipulated in any way. It can also refer to functions, trajectories, or similar elements which can be stored and retrieved from the databases[67].

In addition, Becerra-Fernandez and Sabherwal define data as 'that which is made up of facts, observations, or perceptions which may or may not be correct'. They further argue that data has direct connections with anything that lacks context, meaning or intent, but which can be collected, stored, analysed and distributed, using different media formats[68].

Finally, Du defines data as facts which are recorded to depict, among other things, facts or events which are found in an identified storage medium[69].

Although, as is the case with the FSDoE's Grade 12 results, there may be no meaning attached to the data stored in various media, once DM tools and techniques are applied to such data, valuable information is extracted to help make informed decisions which may directly influence policy making.

2.1.4 Information

Becerra-Fernandez and Sabherwal define information as 'that process of exploitation of raw data in order to identify trends and patterns and make sense out of those output indicators[70]'.

Du further defines information as 'the game of semantics which gives meaning and context to data[71]'. For the purposes of this study, the study applies insights found in the Grade 12 data of the FSDoE that were lying dormant. In addition, such discovered information has helped to identify trends or patterns that contribute to the decision making processes.

2.1.5 Knowledge

Zimmermann defines knowledge as that 'which involves the elements of the mind like being able to comprehend, understand, and learn'. He also quotes Frank Miller (2000) who defines knowledge as:

> the uniquely human capability of making meaning from information—ideally in relationships with other human beings ... . Knowledge is, after all, what we know. And what we know can't be commodified.

---

[67] (Zimmermann 2006)

[68] (Becerra-Fernandez and Sabherwal 2010)

[69] (H. Du 2010)

[70] (Becerra-Fernandez and Sabherwal 2010)

[71] (H. Du 2010)

> Perhaps if we didn't have the word 'Knowledge' and were constrained to say 'what I know', the notion of 'knowledge capture' would be seen for what it is – nonsense[72].

In addition, Becerra-Fernandez and Sabherwal define knowledge by quoting Wing (1999):

> Knowledge consists of truths and beliefs, perspectives and concepts, judgements and expectations, methodologies, and know-how. It is possessed by humans, agents, or other active entities and is used to receive information and to recognize and identify; analyse, interpret, and evaluate; synthesize and decide; plan, implement, monitor, and adapt – that is, to act more, or less intelligently. In other words knowledge is used to determine what a specific situation means and how to handle it[73].

According to Du, knowledge is associated with verified information and structured like 'heuristics, assumptions, associations and models that are understood from data. In other words knowledge adds value to data and information[74]'.

Based on the three interconnected concepts, Du summarises by saying that:

> Data enables an organization to keep records about events that occur. Information enables the organization to react and respond to the events. Knowledge enables the organization to anticipate events and act appropriately when the events occur[75].

For the purposes of this study, the extracted application of DM tools and techniques provides the knowledge that helps to prepare for future occurrences by using the discovered truths as reference points.

2.1.6 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) may be defined as that process which automatically uncovers implicit and valuable patterns from various data reservoirs[76]. KDD can also refer to the entire process responsible for looking for regularity in data and which is done through the application of tools and techniques[77]. Becerra-Fernandez and Sabherwal further define discovery in Databases as 'the process of discovering and interpreting the identified patterns from the data under study through rigorous use of suitable standard algorithms[78]' whereas Du defines KDD as 'a complete process of

[72] (Zimmermann 2006)

[73] (Becerra-Fernandez and Sabherwal 2010)

[74] (H. Du 2010)

[75] (H. Du 2010)

[76] (GARCIA , et al. 2014)

[77] (Giudici 2005)

[78] (Becerra-Fernandez and Sabherwal 2010)

discovering knowledge from data which involves a detailed search for patterns in large unexplored volumes of data[79]'. In conclusion, Pal and Mitra define KDD as 'the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data[80]'.

This study is guided by the KDD which is a body of knowledge that provides a platform for the uncovering of new, trivial and novel information which is made possible through the application of acceptable tools and techniques.

2.1.7 Data Mining

Data mining is:

> the method used for discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques[81].

On the other hand, Giudici defines the term as:

> the process of selection, exploration, and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database[82].

Simply stated, DM refers to the 'act of "mining" knowledge from large amounts of data. Mining is the process that finds a small set of precious nuggets from a great deal of raw material[83]'. According to Refaat, DM is a:

> set of mathematical models and data manipulation techniques that perform functions aimed at the discovery of new knowledge in databases. The functions, or tasks, performed by these techniques can be classified in terms of either the analytical function they entail or their implementation focus[84].

For the purposes of this study, DM tools and techniques were applied to the FSDoE's Grade 12 results data.

---

[79] (H. Du 2010)

[80] (Pal and Mitra. 2004)

[81] (Iranmanesh 2008)

[82] (Giudici 2005)

[83] (Han and Kamber 2001)

[84] (Refaat , Data Preparation for Data Mining Using SAS. 2007)

2.1.8 Educational Data Mining

Educational Data Mining (EDM) is defined as the 'the process where data that is stored in the educational systems is transformed into useful knowledge that helps decision makers to address issues related to the education sector[85]'.

EDM can also be defined as:

> a new field for research which involves the application of data mining techniques on raw data whose origins are in the educational sector in order to address questions and challenges which would lead to the uncovering of hidden valuable information[86].

Educational Data Mining (EDM) can further be referred to as:

> an emerging multidisciplinary research area where the different methods and techniques are used to extract valuable information from the raw data whose origin can be traced from a number of educational information systems[87].

For this study, the data used during the application of DM tools and techniques has its origin in the education field. The questions that are answered are helping the education sector, especially the FSDoE's examinations and assessment directorate to make informed decisions.

2.1.9 Machine Learning

Machine learning is a scientific field which forms 'a sub-discipline of computer science specifically dealing with the design and implementation of learning algorithms[88]'.

Machine learning can also be defined as that which is responsible for identifying any notable relationships and regularities in data which can be transformed into general truths. It can further be referred to as the process which involves the reproduction and data-generation which allows analysts to generalise from the observed data to the new, unobserved cases[89].

For this purposes of this study, machine learning, as a scientific field, has provided a variety of selected algorithms which were automated within DM tools through DM techniques.

---

[85] (Kay, Koprinska and Yacef 2011)

[86] (Pena-Ayala, Educational Data Mining: Applications and Trends. 2014)

[87] (Calders and Pechenizyky 2011)

[88] (Adriaans 1996)

[89] (Giudici 2005)

2.2 Knowledge Discovery in Databases: An Overview

In many studies on DM or KDD, a number of researchers and authors tend to confuse the two terms and, in some cases, treat them as being synonymous. The 1995 Montreal Conference provided a clear distinction between the two concepts. Since then, a number of definitions have been created to give meaning to the two terms. For the purposes of this study, and in compliance with the 1995 Montreal Conference, the two terms are treated separately. Due to a number of approaches by different authors, it is necessary to provide a clear distinction between the two[90].

KDD is the scientific field which focuses on the extraction of raw data of unique and previously-unknown knowledge whose value helps in the improvement of the organizations' daily operations and strategic thinking[91].

In conclusion, and drawing from the Montreal Conference in 1995, an agreement was reached that KDD should refer to the whole process in which information is extracted to create valuable knowledge for the organisation. It is through KDD that relationships between data and the extracted patterns are established in the quest for knowledge[92]. Limiting the study to the DM concept helps to provide a detailed understanding of what it is all about in the whole discovery process.

2.3 Data Mining

Data mining has become a new buzzword in virtually any environment that involves the manipulation or analysis of data. As such, the term has also become overused and misapplied. This means that a clear distinction has to be made between what constitutes DM and what does not[93].The Montreal Conference in 1995 proposed that DM should only be used to refer to that step in which knowledge discovery in the KDD is taking place[94].

The term 'data mining' can be understood literally from the phrase 'to mine' which, in English, means 'to extract'. The verb usually refers to mining operations that extract hidden, precious resources from the earth. The association of the word with data suggests an in-depth search to find additional information which previously went unnoticed in the mass data available[95]. Based on the above

[90] (Adriaans 1996)

[91] (Adriaans 1996)

[92] (Adriaans 1996)

[93] (Adriaans 1996)

[94] (Adriaans 1996)

[95] (Giudici 2005)

explanation, this study plans to extract the hidden patterns from the FSDoE's examinations and assessment database. Like a miner in the bowels of the earth, the study brings to the surface the wealth of interesting patterns which help to expand the body of knowledge.

As the term 'data mining' slowly established itself, it became a synonym for the whole process of extrapolating knowledge. During the process of DM, the main aim is to obtain results that can be measured in terms of their relevance to the owner of the database which would result in a business advantage. Data mining is:

> the process of selection, exploration, and modelling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database[96].

Generally, Data Mining can be described as the process of discovering the hidden knowledge from the company's database which usually leads to the development of patterns whose interpretation gives birth to new rules[97].

The discovered patterns mentioned above must be meaningful in that they must lead to some advantage, usually an economic advantage[98]. The study intends to ascertain whether there are any such patterns in the FSDoE's examinations and assessment database.

Giudici further explains that DM is characterised by the uncovering of patterns and trends to identify opportunities in large databases (such as the FSDoE's examinations and assessment data) for predictive purposes. The tools of discovery involved in this process typically incorporate sophisticated statistical techniques that can be utilised through powerful software packages. These tools are frequently applied to large data repositories, including data warehouses, data marts, and other large data stores[99].

As Figure 2.1 below shows, DM is part of the whole KD process whose responsibility it is to extract knowledge from a variety of databases. The DM process plays a crucial role in applying different tools and techniques which would result in the formulation of different models that would ultimately yield an accurate and acceptable one to be used in future decision making by companies.

---

[96] (Giudici 2005)

[97] (Adriaans 1996)

[98] (Witten and Frank 2000)

[99] (Giudici 2005)

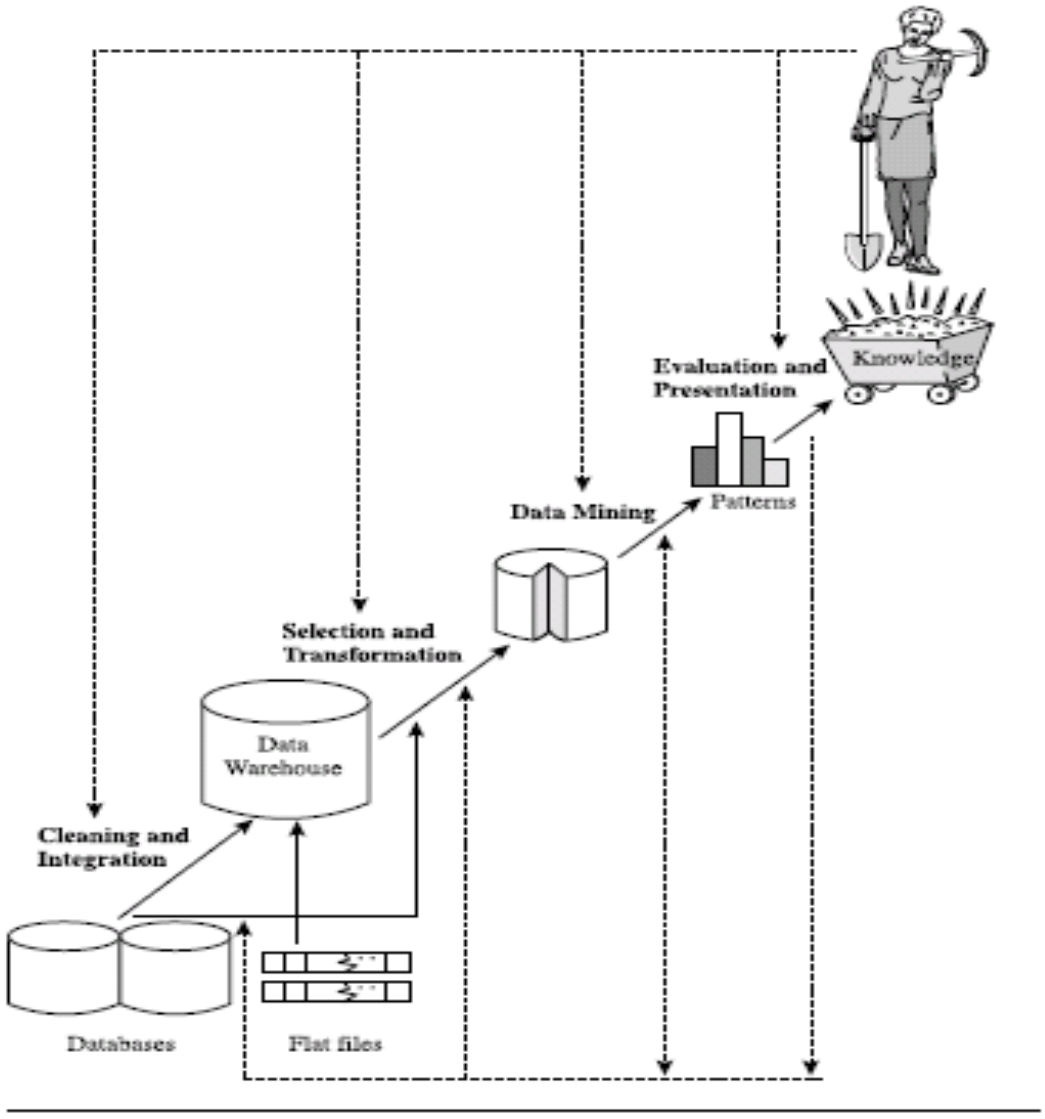


Figure 2.4. 1 Data Mining as part of Knowledge Discovery process (Han & Kamber, 2006:6).

According to Refaat, DM is composed of a set of mathematical models and data manipulation techniques that perform functions aimed at the discovery of new knowledge in databases. These functions or tasks which are performed through these techniques can be classified either in terms of the analytical function they serve or their implementation focus[100].

The important advantage of using DM is the fact that it affords the user or knowledge base an opportunity to engage in direct interaction. The interesting patterns are presented to the user, and may

[100] (Refaat, Data Preparation for Data Mining using SAS. 2007)

be stored as new knowledge in the knowledge base. According to the KDD, data mining is only one step in the entire process, albeit an essential one, since it uncovers hidden patterns for evaluation[101].

It should be noted that a number of companies employ strategies that value the role played by DM as it helps them to understand the data that they generate, and they have realised that it is one of the important drivers of self-learning organisations[102].

The above explanation shows that DM is part of the whole process of extracting knowledge from data, which involves a number of pre-constructed stages that have to be followed in order to achieve the stated objectives. This explanation of DM, as shared by different researchers, helps to ensure that the correct procedures are followed in applying DM to the Free State Department of Education's examinations and assessment database. The data mining stages serve as a 'guiding lighthouse to a wondering ship' which is essential to the successful exploration to uncover knowledge that will help in the decision making process. As indicated earlier, it is of crucial importance that international bodies play an important role in the crafting and development of KDD, to give a detailed explanation of the methodology that guides the process of uncovering new patterns.

2.4 The CRISP-DM Methodology

Many educational institutions have been trying for years to come up with solutions to address, among other challenges, the data with which they are dealing[103]. The area of school performance relies on the manipulation of data in order to make informed decisions. One of the pioneers in this field is W. Edwards Deming, who, in the 1950s, introduced a four-step plan-do-check-act (PDCA) or plan-do-study-adjust (PDSA) cycle aimed at helping to improve educational institutions by using the scientific method of hypothesis, experiment and evaluation. This approach is credited for turning the fortunes of many organisations that were on the brink of extinction[104].

Such pioneering acts led to the creation of a number of methodologies to address the increase in data. In the field of knowledge discovery, there exists a methodology that plays a crucial role in the processing of data[105]. This methodology, which is part of this study, is called Cross-Industry Standard Process for DM or, more concisely, CRISP-DM. CRISP-DM is a step-by-step KD framework within

[101] (Han and Kamber 2001)

[102] (Adriaans 1996)

[103] (V. Bernhardt 2013)

[104] (V. Bernhardt 2013)

[105] (Gilchrist, et al. 2012)

which organisations such as the FSDoE can follow in launching its KD activities. It is regarded as a neutral process model which can be followed in different KD applications.

According to Cios *et al*., the CRISP-DM KD process model shown in Figure 2.2 below is an industrial approach to understanding knowledge discovery[106].

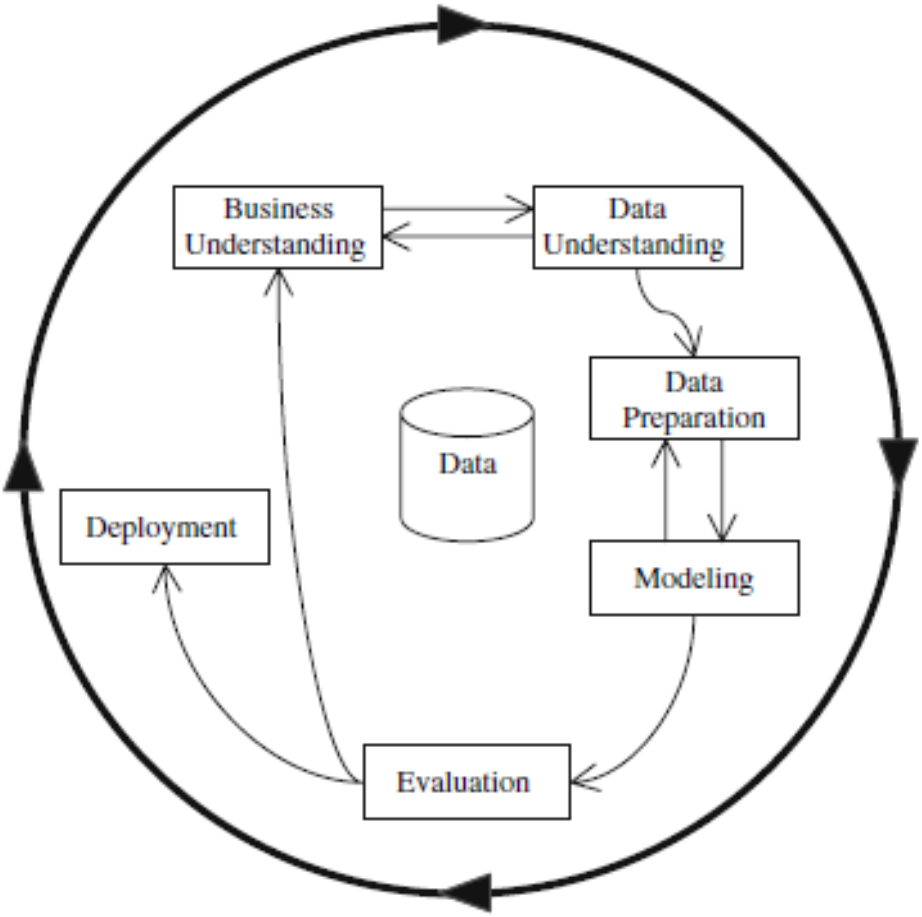


Figure 2.5. 1 The CRISP-DM process model (source: http://www.crisp-dm.org/).

Figure 2.2 above shows the methodological stages that need to be followed in order for the DM process to be successful. The steps that form part of the methodology are business understanding, data understanding, data preparation, data modelling, evaluation and deployment of results.

The graph shown below in Figure 2.3 shows the amount of time spent on the different stages of knowledge discovery. As clearly indicated, and to be witnessed later during the application of DM in the FSDoE's examinations and assessment database, the preparation of data is a stage in which most

[106] (Cios 2007)

time and effort are spent during the KD process. The graph shows that many researchers believe that between 45% and 60% of the time is spent on the data preparation stage[107].

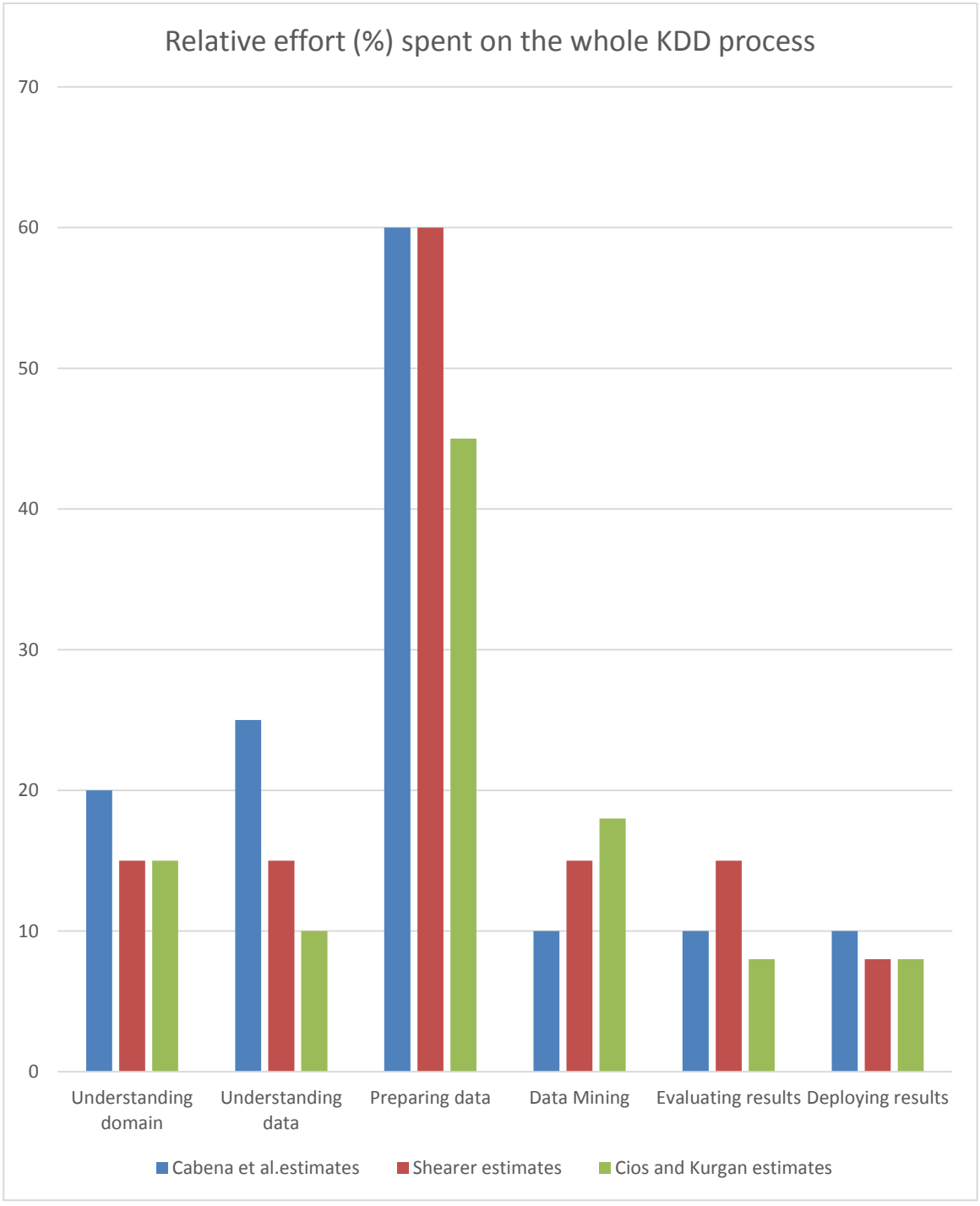


Figure 2.5. 2 Time spent on each step of the Knowledge Discovery process[108].

[107] (Krzysztof , et al. 2007)

[108] (Pal and Jain 2004)

Up until the present day, the CRISP-DM methodology is regarded as a useful model which can work well with a plan-do-check-act (PDCA) cycle in DM applications to obtain optimised quality and success. The eight stages associated with PDCA include problem identification; gathering and selection of data; data preprocessing for missing, duplicate or erroneous information; selection of appropriate learning algorithms; preparation and processing of data; construction and evaluation of the models; interpretation of the discovered knowledge; and, finally, taking action[109].

It is equally important for both the researchers and organisations to have a detailed understanding of the CRISP-DM methodology in order to ensure the successful implementation of the DM tools and techniques. This methodology is not only recognised on a global scale, but also has clear steps that must be followed in order to harvest good results. The CRISP-DM steps are discussed in detail as they have a direct bearing on the study to be conducted on the FSDoE's examinations and assessment database.

2.4.1 Business Understanding

The first requirement on a journey towards the discovery of knowledge is to have a clear understanding of the business problem. This can be achieved by having clearly-stated objectives which would justify the need to execute the DM project[110]. This calls for the creation of an information requirement based on the task to be undertaken which would help to answer basic questions related to the problem. It also assists in making an informed decision about the possibility of applying KDD to the problematic areas[111].

As seen later in this study, the objective is to uncover patterns that form after the application of WEKA's data mining software on the FSDoE's examinations and assessment database.

2.4.2 Data Understanding

Data mining practitioners believe religiously in the saying 'know thy data'. A thorough knowledge of the data with which you are dealing helps to maximise the chances of conducting a successful, efficient and effective knowledge discovery[112]. The data to be used in this study was obtained from the FSDoE's examinations and assessment database. The choice of algorithms or tools is determined

[109] (Guruler, Istanbullu and Karahasan 2010)

[110] (Becerra-Fernandez and Sabherwal 2010)

[111] (Adriaans 1996)

[112] (Becerra-Fernandez and Sabherwal 2010)

by the nature of the database and the type of modifications which suit the DM application requirements. The data is mainly from Grade 12 and represents the marks obtained by learners who wrote their final examinations in the Free State Province. The data spans a four-year period from 2011 to 2013 with 2010 marks to be compared with the other three years. The data represents the actual raw marks which were obtained by individual learners who wrote their Grade 12 examinations in the Free State Province during the focus period of the study. It was obtained courtesy of the Educational Management Information System (EMIS) section which, together with the directorate of examinations and assessment, is responsible for the collection, storage and dissemination of data related to examinations in the province.

2.4.3 Data Preparation

For the successful application of DM tools and techniques, it is important that one spends time on data preparation which involves the following steps: selection and transformation of variables, and data integration and formatting[113]. The collection and selection of data usually takes place after an information requirement has been clearly stipulated[114]. Even though it is not the main requirement, choosing a stable and reliable data warehouse for KDD is always advisable. For this study, assurance was given that the data from the FSDoE's examinations and assessment database represent the original data of the students who wrote their examinations at Grade 12 level.

Selecting data begins by identifying a database which contains the raw data to be mined. The chosen data is copied from the database and stored separately[115]. When it comes to the FSDoE's examinations and assessment database, a sample which represents the entire dataset has been selected. This dataset, as indicated earlier in the limitations section, forms part of this study from the selected Free State schools in the Motheo District.

As in any DM applications, and as is the case with the FSDoE's examinations and assessment database, some of the dataset has been re-constructed and transformed in order to meet the requirements of the selected DM software as without it, it would have been difficult to produce effective models. The collected data was cleaned and, in cases where the database is large, taking a sample is always a good option. One common aspect when dealing with databases is that they always have anomalies which may negatively impact the entire DM process. It is important that the process

[113] (Becerra-Fernandez and Sabherwal 2010)

[114] (Adriaans 1996)

[115] (Adriaans 1996)

of data cleaning is followed in detail as doing so during this initial stage saves a great deal of time and money[116].

The cleaning of data is done for various reasons, including, among others, the duplication of records, the correction of typing and spelling errors, and so on. Given some of the similarities between the two, data cleaning can easily be mistaken for data mining[117].

The enrichment of data follows once data cleaning has been completed, and entails obtaining additional information which may be bought from private organisations dealing with a variety of subjects, including the problem that is being addressed[118]. Sometimes data enrichment is achieved by seeking free additional data or having a verbal conversation with specific people to enrich the data undergo data mining[119].Data coding during DM preparation involves making conscious decisions to either overlook or delete some of the data if it is deemed to have no value in terms of the process[120]. As is the case with the FSDoE's examinations and assessment database, some datasets have undergone the process of coding in order for them to be accepted for further data mining processes. Data coding may also be achieved by removing the noise in the data through the simple method of filtering out the affected records[121]. The data that has been used from the FSDoE's examinations and assessment database has been cleaned, enriched and coded in order to meet the objectives of the study which are to seek patterns in the Grade 12 results as observed over a period of time.

This is done as an answer to the requirements of the WEKA software which requires reformatting before any application can take place on the FSDoE's examinations and assessment database. This step involves re-ordering and reformatting of the data fields as required by the DM Model. In this study, data is formatted into ARFF or csv in order to meet WEKA requirements.

2.4.4 Data mining

This stage, which forms the backbone of this study, is thoroughly discussed in the following section. The FSDoE's examinations and assessment database has been subjected to DM tools and techniques

[116] (Adriaans 1996)

[117] (Adriaans 1996)

[118] (Adriaans 1996)

[119] (Adriaans 1996)

[120] (Adriaans 1996)

[121] (Adriaans 1996)

which help to determine whether there are patterns to be identified in the Grade 12 raw data as observed over a selected period of time.

2.4.5 Evaluation and Interpretation

Once the model has been determined (as is the case in this study of the FSDoE's examinations and assessment), the validation data set is compared with the actual results. This comparison helps to ensure that an accurate model is confirmed[122]. The WEKA software, which has been used in this study, evaluates various models that have been generated from the FSDoE's examinations and assessment database. The application's results help to interpret the generated models in relation to the research question posed in the first chapter.

2.5 Comparison between KDD and DM

Although the two terms should not be viewed as synonymous, there is a great deal of evidence that points to the existence of a relationship. DM is one of the important steps in the KDD which is responsible for the practical uncovering of valuable information and knowledge from data[123].

In comparison with the KDD process, Data Mining serves as the mathematical core which involves inferring algorithms that explore the data, developing mathematical models and discovering significant patterns (implicit or explicit) – which are the essence of useful knowledge[124]. The pattern in the given data is an expression that helps to interpret the data or a model applicable to the subset in the given data[125].

Knowledge Discovery in Databases (KDD) is generally made up of six steps, namely data selection, cleaning, enrichment, coding, data mining and reporting. Of all the stages, DM represents the actual discovery of knowledge from the raw data. As an on-going and flexible process, KDD cannot be accomplished by following the steps one after the other. During any step in the process, a data miner could step back or move forward to address whatever would need to be taken care of at that moment. This means that for organisations to earn the title of being a learning institution, they must allow DM to take place on an open-ended platform where data fixing is performed as an ongoing exercise[126].

---

[122] (Becerra-Fernandez and Sabherwal 2010)

[123] (Pal and Jain 2004)

[124] (Khandar and Dani 2010)

[125] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

[126] (Adriaans 1996)

Another interesting aspect of the stages involved in the KD/DM process is that they do not have to be followed religiously as they allow for flexibility in terms of moving backwards or jumping forwards to any stage, anytime.

In conclusion, and drawing from the Montreal Conference in 1995, an agreement was reached that KDD should refer to the entire process in which information is extracted to create valuable knowledge for the organisation. It is through KDD that relationships between data and the extracted patterns are established in a quest for knowledge, and DM forms part of the different stages that make that possible[127]. The detailed explanation of KDD and DM has created a fertile ground for understanding what this study intends to accomplish. Basing on the above information, an explanation of Educational Data Mining (EDM) helps to highlight the role played by the application of DM tools and techniques in an educational setting.

2.6 Educational Data Mining

Linked to this study on the FSDoE's examinations and assessment database is the understanding of the role played by educational data mining or EDM in the broader field of data mining. Despite being overwhelmed by data, it has been observed that teachers spend little time on analyzing factual data pertaining to students and test items. The development of computer technologies such as EDM promises to be a solution to the challenge of analyzing and interpreting the raw data generated in classrooms on a daily basis[128].

Educational Mining (EM) as EDM is sometimes called, is concerned with exploring educational data (ED), using a variety of data mining techniques. As seen later in this study about the FSDoE's examinations and assessment database, EDM also makes use of a number of statistical and machine learning approaches to understand data in the education environment[129].

Despite being a new field of study, Educational Data Mining prides itself in being a data mining solution to all matters pertaining to the exploration of data from an educational environment perspective. It focuses more on exploiting data from an educational perspective and developing descriptive patterns, and predicts many important elements relating to learners, among others, learner behaviour, assessment and achievement. A number of approaches to EDM concentrate on many characteristics that have a bearing on students' learning such as acquired domain knowledge,

[127] (Adriaans 1996)

[128] (Ivancevic, Knezevic and Pusic 2014)

[129] (Srimani and Patil 2012)

personality, and academic achievements which are explored by using a number of machine learning methods. Pena-Ayala also argues that Educational Data Mining explores a number of trends in an educational setting such as text mining and social networks analysis[130].

What also causes this new data-oriented technology called EDM to stand out is its ability to uncover valuable information as the result of an interaction between humans and computer technologies[131]. In addition, EDM transcends different fields of study such as computer science, education, statistics learning analytics and machine learning. The choice of method or technique to be applied during the EDM depends on the type of educational challenge to be addressed[132].

It is through direct interaction with EDM that learners, teachers, researchers and administrators gain a great deal from the activities that the technology provides. Learners, for example, are assisted in paving the way that leads them to success. Teachers, on the other hand, are given a unique opportunity to experiment with a number of teaching strategies that allow them to improve the way in which they relay the learning process to their students in the classroom. Researchers are able to conduct a comparative analysis of the output results obtained after the application of various DM techniques on the targeted data. Lastly, administrators are able to formulate strategies that would assist in the improvement of various educational entities that fall under their ambit[133].

A number of countries are acknowledging the role that data plays in improving educational processes. Trends can also help the education sector to put measures in place that lead to the improvement of the entire system[134].

In a school environment, data analysis may help the school to obtain a clear picture of how it has been performing in the past, and to identify the current professional learning and teaching challenges that have to be met in order to create a better future for its clients. Data assists teachers in achieving a clear understanding of how their students learn, and providing the necessary support that they need. Proper utilisation of the educational data informs decision makers not only about human characteristics but also about how the different parts in the education sector relate to one another to form a complete whole[135].

---

[130] (Pena-Ayala, Educational Data Mining: Applications and Trends. 2014)

[131] (H. a. Guruler 2014)

[132] (Bousbia and Belamri 2014)

[133] (Bousbia and Belamri 2014)

[134] (V. Bernhardt 2013)

[135] (V. Bernhardt 2013)

Demographic data, for example, is widely used in an educational environment due to its dynamic nature. It can give researchers an idea of, among other things, a school's philosophy through indicators such as how students are disciplined and identified for special education, advanced placement, gifted programmes, etc.[136]. The student learning data, for example, can tell whether the selected students meet the requirements by uncovering their strengths with regard to learning, and identify areas that require improvement. By examining student learning data, EDM can show whether a school has managed to achieve a continuum of learning for students, established instructional coherence, alignment of the curriculum, and instruction and assessment within and across grade levels[137].

It is critically important to conduct real-time interventions for those students whose performances do not meet expectations, and this includes those who risk being labeled as poor performers. It is a known fact that students' academic performance has a bearing on the smooth integration into academic life and therefore it is the responsibility of teachers to plan ahead and prepare students' activities based on their levels of development. Students' modelling techniques can employ EDM by delving more into the field of study before assessing their knowledge of it. It is through the development of such models that learners' knowledge can be assessed. As it is impossible to measure knowledge, performance-measuring tools such as the Bayesian network, for example, may be used to model learning processes based on Markov's assumption that current knowledge is directly linked to prior knowledge and the learning activities[138].

Garcia *et al*. believe that a normal educational data mining process can be accomplished by following the steps listed in Figure 2.7.1 below. The educational environment is at the centre of the data mining study, and the results are applied to it to remedy what was observed as a challenge[139].

[136] (V. Bernhardt 2013)

[137] (V. Bernhardt 2013)

[138] (Moradi, Moradi and Kashani 2014)

[139] (Garcia, et al. 2011)

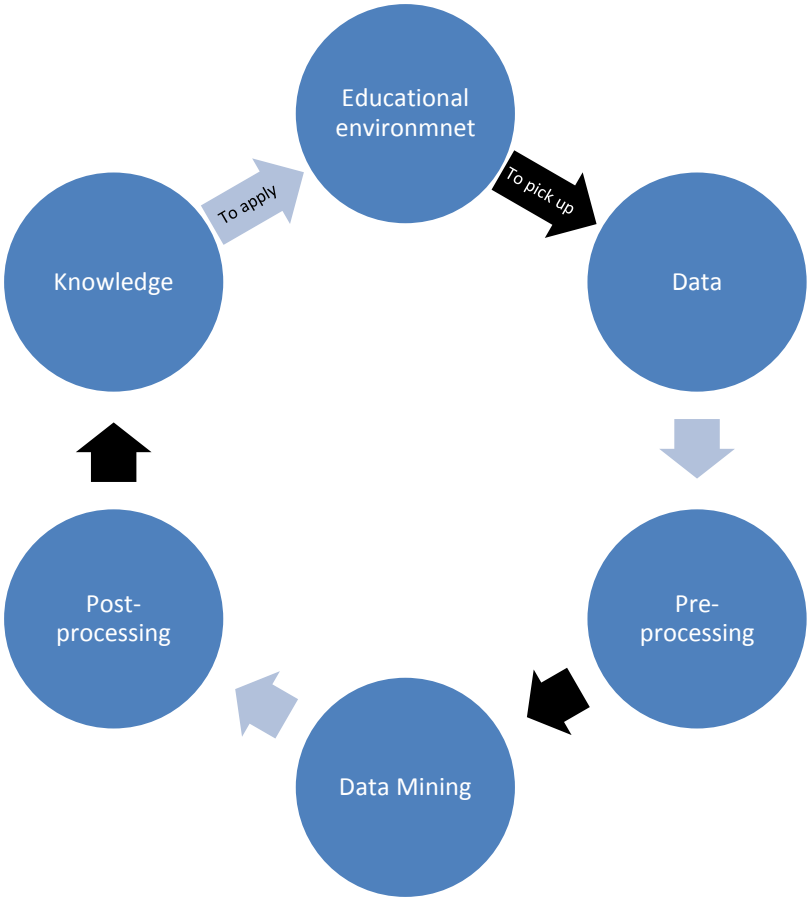


Figure 2.7. 1 Educational Data Mining process[140].

EDM technologies have a framework that is developed and applied by copying from a universally-accepted methodology called CRISP-DM. The availability of such a methodology means that an EDM study has to follow the steps that have already been set up until the desired results are obtained. A successful EDM exercise rests heavily on the use of a number of DM techniques such as classification, clustering and association analysis that have proved to be popular with researchers in the field[141].

Despite being new in the field, EDM has caught the eye of many in the education sector. The entire world is either fully embracing or slowly moving towards the use of EDM to find data-related solutions in the education sector.

[140] (Garcia, et al. 2011)

[141] (Bousbia and Belamri 2014)

2.7 The Application of EDM on a Global Scale

Even though it is still an emerging field of study, the practical application of EDM has produced good results in the United States who are the leaders and who are far ahead of European countries and the rest of the world[142].

The success of EDM in many countries depends on the governments' willingness to embrace the new technology and offer support at all levels. In an ideal world, elements such as educational tests, skill assessment, mastery of competencies, and students' ratings are the determinants of the quality of education. Quality relies heavily on the use of a reliable computer information system which helps in the execution and production of helpful results from data understudy. On a global level, EDM helps decision makers at all levels in the education sector to have a better understanding of the data that is produced on a daily basis. Notable areas that can benefit include the methods employed during the teaching of learners, offering hints to students when it comes to problem solving, and assessing the various subjects taken by learners[143].

In 2001, for example, in response to the challenge of the educational data scattered all over, the United States crafted a data-driven policy that was to be implemented by the education sector. This top-level decision led to the development of the state-of-the-art data warehouses across the United States. They realised that if they continue to ignore the role played by data in the education sector, they risk finding themselves drowning in it in years to come without any formula in place to help them understand it. A number of systems were created to address the challenges posed by the data overflow[144].

The well-known systems that were developed in the US include the Achievement Reporting and Innovation System (ARIS) in New York in 2007, an Integrated Resource Information System (IRIS) in Milwaukee in 2008, and Longitudinal Educational Analytics and Decision Support System (LEADS) in Nashville[145].

The mushrooming of these data warehouses led to an interest in data itself. One of the largest and most important research programmes on educational data in practice originated in 2006 when the DoED launched a study on *Education Data Systems and Decision Making* which was conducted by

[142] (Adriaans 1996)

[143] (Pena-Ayala and Cardenas, How educational data mining empowers state policies to reform 2014)

[144] (Piety 2013)

[145] (Piety 2013)

the Centre for Technology and Learning at SRI International, a research organisation that has worked extensively with education technology[146].

Many countries, such as Mexico, have realised the importance of involving learners when it comes to making decisions about their future. Through EDM, they are able to obtain valuable information from a number of surveys that they conduct involving learners. It is through such exercises that the government is able to obtain diverse opinions which help to pave the way towards quality education[147].

In many educational institutions of higher learning, KDD has proven to have a positive influence on the evaluation of student data. It is through the analysis and manipulation of data that students gain satisfaction from the knowledge that they are offered quality education. The emergence of a variety of technologies on data manipulation helps institutions of higher education to improve the efficiency and effectiveness of their traditional processes. Many institutions are developing models to analyse students' results over a period of time in an academic year[148].

Many institutions of higher learning, for example, rely on a Student Relationship Management system (SRM) to bring together all the data relating to a student's experiences. SRM can then be manipulated to predict the student's progress which may help those who are struggling to be placed in a remedial assistance programs[149].

Conducting EDM has become a popular exercise in many institutions of higher learning. In one study, for example, it was used to predict academic success. The main objective was to predict, among other things, dropouts early on in a period of study, graduation on time, general performance, or the need for remedial classes. What is unique to this study is that the datasets, which were collected from the entire university, were not only large but were also collected over a period of time. A number of classification methods were compared to each other, for example, the decision trees, Bayesian networks and neural networks, respectively. A 79% accuracy rate was obtained and that depict the success of the project[150].

In another study, in which EDM was used to predict course outcomes, the main focus was on predicting a pass or a fail, dropouts and test score. A variety of classification methods was applied

[146] (Piety 2013)

[147] (Pena-Ayala and Cardenas, How educational data mining empowers state policies to reform 2014)

[148] (H. a. Guruler 2014)

[149] (H. a. Guruler 2014)

[150] (Hamalainen and Vinni 2011)

iteratively with the decision trees, Bayesian Networks, neural networks, K-nearest neighbour classifiers and regression-based methods, all of which played a central role, respectively. The average accuracy was 72%, with the best cases at 90%[151].

Another experiment was conducted to classify meta-cognitive skills and other factors which affect learning. The study was used to predict, among other things, students' motivation or engagement level, cognitive styles, expertise in using the learning system, 'gamming' the system, or recommended intervention strategy. The most commonly-used classification methods include decision trees, k-nearest neighbour classifiers and regression-based techniques. Classification accuracy varied between 88% and 98%[152].

The above practical applications of EDM are a testimony to the fact that when the technology is properly applied to data in education, the outcomes could contribute towards sustaining better, quality-driven education. The important aspect that many researchers cite is that EDM assists organisations in conducting a thorough investigation and choosing the correct technology to manage and understand their data. It is by acquiring the best EDM technology that organisations are able to benefit from the data that they create on a daily basis.

2.8 Benefits of establishing a Knowledge Discovery System

Given the well-constructed and well-defined stages that need to be followed during the DM process, many questions are always asked regarding the benefits of having a knowledge discovery system in a company.

The successful application of data mining technologies by institutions such as the FSDoE to their examinations and assessment database requires the selected knowledge discovery system to fit the strategy of the institution. The knowledge discovery system must be able to answer data-related questions. Fayyad and Uthurusamy cleverly ask 'now that we have gathered so much data, what do we do with it?' According to them, this question has become common in many organisations in that they have realised that the presence of proper digital technologies has made the capturing and understanding of data an unavoidable activity[153].

[151] (Hamalainen and Vinni 2011)

[152] (Hamalainen and Vinni 2011)

[153] (Fayyad and Uthurusamy, Data mining and knowledge discovery in databases. 1996)

Fayyad and Uthurusamy further argue that the true value of data lies in the ability to extract useful information whose results directly influence future decision making[154]. Maimon and Rokach further believe that knowledge discovery demonstrates intelligent computing at its best, and is the most desirable and interesting end-product of Information Technology. Discovering and extracting knowledge from data is a task that many researchers and practitioners wish to accomplish. This study on the FSDoE's examinations and assessment database was prompted by the above arguments which led the researcher to believe that there is a great deal of hidden knowledge waiting to be discovered[155].

Qin further states that the benefits of having a knowledge discovery system can only accrue if there is clarification with regard to the kinds of knowledge that are to be discovered as this helps one to decide on the type of database that one needs to work on and the techniques to be used for discovering the anticipated knowledge. He believes that this is the first and most important step in KDD[156].

2.9 Conclusion

In conclusion, the success of any DM activity depends on following the universally-accepted CRISP-DM methodology whose advantage is that it can be applied in an interactive and iterative manner. The abundance of literature on the DM application presents a variety views and critical analyses, all of which help to provide a better understanding of the field.

The field further calls for the exploration of literature on databases. A number of steps are involved in the development, choice and application of databases on the datasets. Obtaining a number of views with regard to databases is of assistance during a later stage as the study applies various tools and techniques on the FSDoE's examinations and assessment database.

[154] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

[155] (Fayyad and Uthurusamy, Data mining and knowledge discovery in databases. 1996)

[156] (Qin 1999)

Chapter 3 *Databases as data source*

On a daily basis, organisations generate data which is usually stored in their databases. Though unexplored, such data has the potential to generate valuable information provided that proper technologies are used to unlock it[157]. Understanding data sources assists in knowing more about the nature and character of data that organisations have in storage.

This overwhelming data creates a situation which can be described as being 'data rich but information poor'. It may also mean that data generation continues to exceed our ability to understand it. In order to solve this challenge, many organisations have established databases which are filled with what they refer to as 'data tombs'. Many of these data archives remain dormant for years with no one paying attention to them. Abandoning such data results in companies making uninformed decisions based on intangible evidence. Such decisions emanate from the unavailability of suitable tools to extract valuable knowledge lying deep in the raw data found in the databases[158]. Decision making relies heavily on the existence of reliable data found in databases. Such reliable databases then assist researchers in extracting and appropriately analysing data, to make them meaningful and useful[159]. In order for the successful exploration of data to occur, it is important to delve into how databases, as data sources, have been evolving over the years.

3.1 The historical evolution

The history of data storage devices and their usage can be traced back to the 1960s when Charles Bachman coined and designed the first database management system[160]. This integrated data store as it was called at the time was endorsed by the conference on data systems (CODASYL) as a foundation for designing and understanding data sources. Based on Bachman's seminal work, IBM created an Information Management System (IMS), DBMS, in the late 1960s which is not only still used today but was expanded into a new hierarchical data model[161].

As a well-known founder of the Integrated Data Store, Bachman, a winner of the 1973 ACM Turing Award, delivered a seminal speech entitled 'The Programmer as Navigator'. He argued that many

[157] (Adriaans 1996)

[158] (Han and Kamber 2001)

[159] (Chowdhury 2009)

[160] (McJones 2009)

[161] (McJones 2009)

people believed that data has the ability to flow through a computer, through a program, in a sequential form from one end to the other. He suggested that computers should, instead, be seen as a platform where data records move through and are connected to each other by links[162].Bachman's ideas attracted a great deal of interest, and the field of study began to expand.

The 1970s gave birth to a new data representation framework called the relational data model which was founded by IBM employee and well-known Mathematician, Edgar Codd. Codd, who won the 1981 Turing Prize for revolutionising the DBMS, developed a relational model which is associated with many database management systems that are in still in use by many organisations today. Despite their popularity and being the DBMS of choice in almost all organisations, relational models played a major role in the advancement of scholarly work and research on the role played by database management systems in many organisations[163].

The late 1980s led to the further development of structured query language (SQL) by IBM whose focus was on relational databases. The SQL was given the stamp of approval and was standardised by well-known international bodies such as the American National Standards Institute (ANSI) and the Institute for Standardization (ISO)[164].

In the 1990s, a number of developments, whose focus was on database systems and advanced query languages, took place. A number of complicated and richer data models were created in response to a widening variety of data that had to be stored by organisations. Chief among the new forms were those in image format. Such modified database systems were not only able to deal with complex data queries but could bring together data from various databases to perform unique and specialised data analysis. Well-known database systems such as Oracle and PeopleSoft have the ability to consolidate data from a number of databases within an organisation and provide a platform for various data-related tasks to be carried out[165].

In this age of the Internet, many organisations' websites were used to store data in operating systems, whereas they now make use of the DBMS to store their data through a web browser. In this way, queries are formulated in a web-based format and the responses are formatted through the use of mark-up languages such as HTML[166].

---

[162] (McJones 2009)

[163] (Ramakrishnan and Gehrke 2003.)

[164] (Ramakrishnan and Gehrke 2003.)

[165] (Ramakrishnan and Gehrke 2003.)

[166] (Ramakrishnan and Gehrke 2003.)

Databases vary in design and the functions they support. Organisations have operational databases which focus mainly on allowing access to their daily transactions. Some have data warehouses which are integrated data stores that are built from the operational databases. Their usage is largely strategic in nature and provides decision makers with information that can help to sustain the future of an organisation[167]. Their uniqueness and the type of data that they store can help to offer solutions to many of the challenges faced by many organisations today.

According to Campbell, it was in 1996 and 1997, respectively, that a chess computer called Deep Blue managed, for the first time, to defeat Garry Kasparov, a human world chess champion. The success of Deep Blue was attributed to its ability to access and utilise valuable knowledge that was stored in 700 000 databases. Drawing on Deep Blue's story, it is safe to say that databases provide storage platforms for different types of data which, as seen later in the study on the FSDoE's examinations and assessment database, helps organisations to draw conclusions based on the models that are generated during the application of DM tools and techniques[168]. Deep Blue succeeded because it had data that was not only available in abundance in databases, but also relevant in terms of answering all the challenges faced by the machine.

In addition to the above, a recent survey revealed that databases are still considered as being valuable knowledge management tools recommended by organisations in the 21st century[169]. Antonova (2011) found that 60% of 357 managers consider databases as valuable tools when it comes to knowledge sharing[170]. For businesses or organisations, databases represent a critical knowledge management tool that must be leveraged to yield maximum strategic returns. Nowadays, having knowledge and utilising it in organisations has become a process which can give them an advantage over the competition[171]. A thorough understanding of databases is important as their structure and operating procedures play an important role in convincing decision makers that whatever results are produced by them are reliable and credible.

3.2 Overview of Database systems

In order to gain a thorough understanding of database management systems (DBMS), it is important to know how the data hierarchy is formed. Data items form the basic component of a file. In a file

[167] (Adriaans 1996)

[168] (Campbell 1999)

[169] (Gilchrist, et al. 2012)

[170] (Gilchrist, et al. 2012)

[171] (Gilchrist, et al. 2012)

system, a combination of a number of related data items results in the formation of a record[172]. In addition, a number of similar records put together results in the formation of a file. In relational theory, data items are referred to as columns or attributes, whereas records are referred to as tuples. It is through the use of DBMS, software designed to manipulate databases, that we are able to navigate databases with ease[173].

The role of Database Management Systems (DBMSs) is to give identity and meaning to the organisation's data. This data, which is found in various databases in many organisations, can, through the use of DBMSs, help to uncover the nature and characteristics of the stored data, and point to the existence of relationships between entities[174]. Many DBMSs in use today are based on the well-known relational model which briefly gives users the ability to define and retrieve the data that they want without having to worry about the complicated processes involved that would make such a request possible[175].

The main duty of the DBMS is to ensure that the selected data is grouped, stored and later accessed in the databases without any difficulty. The DBMS forms a link between what the user wants and what the databases have and can provide. It is the databases that form a storeroom for all the created tables. As already mentioned, many DBMSs follow a relational model which allows for the creation of relationships in data within a database and, in many cases, link with neighbouring databases[176].

The function of a database is to help integrate the organisation's scattered data into a common and easily-accessible source. A simple database is composed of entities which, for example, in an institution of higher education, would be a faculty, a department, a lecturer, a course and students who take that specific course. Furthermore, databases can provide a link between entities, for example, in a faculty, a certain lecturer may be teaching a course to a certain number of students. In circumstances where organisations have to deal with enormous databases, DBMSs are available to navigate through such tons of data in order to help managers and general users make informed decisions. One of the well-known and popularly used DBMSs is a relational database system[177].

---

[172] (Teorey, et al. 2011)

[173] (Teorey, et al. 2011)

[174] (Ramakrishnan and Gehrke 2003.)

[175] (Ramakrishnan and Gehrke 2003.)

[176] (Norman 2004)

[177] (Ramakrishnan and Gehrke 2003.)

Drawing from the earlier definitions of data, which is simply described as observable facts, a database provides a platform for such data in which to be stored. Databases also allow for the manipulation of such data which normally leads to the creation of different models that focus on areas such as in the education sector. The abundance of electronic databases allows for the creation of a number of models which can help all interested parties to make informed decisions[178].

3.3 The Relational Databases

It is worth noting that before the advent of relational databases, the retrieval of information was difficult in the sense that it required deep knowledge of how computer programs work and, specifically, knowledge of how data can be stored and accessed. Such an exercise was not only time-consuming but also costly. It was only in the 1970s that an Oxford graduate named Tedd Codd ventured into the study on how information that is stored in large databases can be retrieved by the end-user who would not have to worry about how it was stored or structured[179].

It is important to note that the term 'relation' has a strong association with the subject of Mathematics which addresses the idea of the existence of a logic or relationship among objects. The relational database theory also has its origin in the concept of relation which addresses, among other components, relational algebra, variables and operators[180].

Relational databases, which were pioneered by Tedd Codd, revolutionised the way in which business data was handled and understood. His well-known 1970 publication on the nature of relational data management entitled 'A Relational Model of Data for Large Shared Data Banks,' resulted in the development of new and trusted ways of interacting with and manipulating data in computers[181]. In this innovative way, Codd insisted that data should be viewed at as an independent entity and that end-users should concentrate only on their specialised business and less on how the data was stored and accessed by computers[182].

Codd further argued that the interaction between the end-users of data banks should be less concerned about how the data is internally structured, and more about retrieving and accessing specific information to meet a particular need. In addition, while being cognizant of the need to upgrade

[178] (Dringus 2005)

[179] (IBM n.d.)

[180] (Darwen, an-introduction-to-relational-database-theory-ebook 2014)

[181] (Smith 2005)

[182] (IBM n.d.)

application programs, he contends that such changes should not affect the daily operations of end-users. He predicted that 'Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information[183]'. Codd made use of the concept of relations which was not only widely used in Mathematics but also contains elements that could be applied in data studies.

As previously alluded to, the term 'relation' has a strong association with Mathematics. Therefore, it is able to identify and show the existence of relationships between objects in a domain. It is common, for example, for businesses to look for a relationship between the products that they sell and the customers who consume them. They would, in a mathematical relation be able to find a link between the products and the customers[184]. Codd wanted to apply the same mathematical principle in a way that would allow users to interact with stored data and then receive the independently-created results which would give out relationships to the user. He referred to this process as data independence. He was opposed to the use of pointers, indexes or any procedural activities to carry out searches. He wanted users to imagine the existence of an abstract relationship between objects which can be defined along the same lines as between products and customers[185].

The use of abstract relations in stored data was complemented by the introduction of a number of languages whose role it was to manipulate the stored data. One such language was relational calculus which is based on logical notation in calculus which permits one to ask questions through the process of binding variables. Such abstraction in querying data was made to function independently of the container in which data was stored[186]. Another language that Codd used was called Relational Algebra which tended to be procedural in nature in the sense that it is made up of a number of operators which would bring relations between objects together[187]. He believed that the language would be complete if it met algebra's expressive strengths[188]. Relational algebra's several operators are usually expressed in the form of symbols. A relational operator is known for taking one or more relations as operands to produce a relation. Such operators form a basic set that is good enough to represent a complete

[183] (Codd 1970)

[184] (McJones 2009)

[185] (McJones 2009)

[186] (McJones 2009)

[187] (McJones 2009)

[188] (Darwen, The Relational Model: Beginning of an Era. 2012)

relation. It is always said that 'a language is deemed relationally complete if it supports directly or indirectly, all of the operators of 'that' algebra[189].

The main role players in the creation of a data model are a designer and the client who influence changes throughout the developmental stages. A well-crafted data model has the ability to communicate clearly the correct message to the user. In a successfully-designed data model, the attributes, entities, identifiers and relationships play an important role which satisfies the original goals set out by the client[190].

A typical data model is made up of three components, namely data structure, integrity and operators, which are critically important role players during the accessing, modification and expansion of data. To manipulate data in the database, there are a number of languages available which are used, including relational algebra. Relational algebra is recognised as a standard that assists in the evaluation of the data retrieval languages. The eight available relational algebra operators which are responsible for the placement of data queries connect the existing relations or help in the creation of the new ones[191].

Finally, in the mid-1980s, Codd came up with twelve commandments which are applicable to relational database systems. It is through these commandments that Codd wanted to provide a clearly-defined framework for the way in which data in the database is to be managed[192]. A thorough understanding of the language used help to know how data sources are explored in order to retrieve data.

3.4 The role of Structured Query Language (SQL) in databases

SQL is a popular language that is associated with the creation, manipulation, and querying of relational DBMSs[193]. Users are able to receive datasets that they require from databases by formulating queries that are converted into a language that is only understood by a DBMS. The formulation of such technical language is made possible through the utilisation of certain tools called Structured Query Language (SQL). Such tools have the ability to carry out complicated logical

[189] (Darwen, an-introduction-to-relational-database-theory-ebook 2014)

[190] (Watson 2006)

[191] (Watson 2006)

[192] (Watson 2006)

[193] (Ramakrishnan and Gehrke 2003.)

processes which may involve the formulation of summaries, sorting and joining together a variety of tables in order to help the user retrieve the data they requested[194].

To accommodate the evolving SQL technology, a number of features from outer joins, table expressions to Online analytic processing (OLAP) functions were added. Despite SQL being criticised for moving away from Codd's foundation, it has been observed that the language is 'simple enough to learn easily and expressive enough to do useful work[195]'. In the field of databases, SQL plays three prominent roles which include data definition language (DDL), data manipulation language (DML), and data control language (DCL). Although it is not a programming language, SQL works hand-in-hand with well-known programming languages such as Java to create application programs[196]. An increase in the interest in database technology led to the creation of standards that will have to be adhered to by those involved in the field. A number of standards have been developed over the years, the latest of which is SQL 2003 and SQL 2006. It is through an SQL interpreter that a query is placed which may include, among others, a request for specific data, for the deleting of data or the creation of new meta-data. It is the SQL that forms a link between a request and the DBMS[197].

According to Chamberlain, it was impressive to see Codd's use of compact relational language to address complex queries. Such developments also made it possible for designers to come up with a relational language, the content of which would be easily understood by people who do not have a background in either mathematics or computer programming[198]. Despite its success, Codd's language had challenges in that its mathematical notation proved to be difficult to enter on a keyboard. Overcoming this challenge required replacing symbols with keywords. Secondly, his adopted language bore resemblance to mathematics' set theory and symbolic logic, and many commentators called for the use of a language which would cut across disciplines and be understood by everyone[199].

3.5 Data Warehouses

In the mid-1990s, Kimball played an important role in the creation of dimensional modelling which led to the creation of advanced data warehouses, many of which are still in use by many organisations

[194] (Provost and Fawcett 2013)

[195] (Chamberlain 2012)

[196] (Watson 2006)

[197] (Norman 2004)

[198] (Chamberlain 2012)

[199] (Chamberlain 2012)

today. Unlike the operational systems whose main concern is daily operations, data warehouses are involved in the evaluation of the organisation's processes[200].

One of the notable functions of data warehouses is to bring together and consolidate data from diverse transaction processing sources within organisations. It is common to discover that the individual transaction systems that are scattered throughout the organisation have their own unique databases. Through their technology, data warehouses make it possible for data-analysing systems to carry out data mining in order to produce models which help organisations to make informed decisions[201].

A successful acquisition and installation of a data warehouse in an organisation depends on its ability to satisfy the needs of the users across all departments. It must also be born in mind that there is no one-size-fits-all approach when it comes to storing, modifying, and accessing data from the data warehouse. Kimball and Inmon played an important role in laying down a foundation and principles for acquiring and using data warehouses by organisations[202].

According to Kimball, data modelling allows for the data to be structured in a logical way that makes it easier for users to place data queries and then receive the desired results within a short period of time. This modelling technique is widely known for its simplicity and effective data warehousing presentation abilities[203].

Architecturally, data warehouses have always been at the forefront of information systems, and play a vitally important supporting role in the established integrated, historical data. They organise, and store the data which is later accessed, retrieved, and analysed, and assist in making informed decisions. As previously mentioned in earlier chapters, data warehouses are subject-oriented, integrated, time variant and non-volatile in nature. The daily operations of an organisation is usually the source of the data that is stored in data warehouses, and the two environments are separate from one another[204].

Data flows in a seamless way and undergoes a number of transformations as it passes the different levels mentioned above. It starts from the operational process and ends up in the aging process. The Data that is mostly used is that which resides at the summarisation level compared with that which is

[200] (Corr and Stagnitto 2013)

[201] (Provost and Fawcett 2013)

[202] (Silvers 2008)

[203] (Kimball, et al. 2008)

[204] (Inmon 2000)

at the old or aging level[205]. Furthermore, it is important to note that data can be considered as being static or dynamic. An operational system is dynamic whereas a data warehouse is static. Unlike static data, dynamic data is an ever-changing mountain of data which cannot be easily explored to provide the desired results. This leads to the conclusion that data stored in databases appears to be structured, whereas that in data warehouses is loosely structured[206].

3.6 Conclusion

In the present knowledge economy, quality forms the backbone of all the activities that take place in organisations. It is incumbent upon all decision makers to ensure that data quality in the educational outputs can be both improved and validated. For example, having quality-laden database systems in the education sector is more than enough to convince various stakeholders that the output levels are not only credible but also of a higher quality. Databases are sources of knowledge which is an end-product of data exploration, using DM technologies.

Based on the above, many organisations have realised that data creation is increasing and that they must always be on the lookout for suitable data sources which can be used to store and later explore such data to their advantage. In order for organisations to access and make use of correct and valuable information, they need to invest in reliable databases which are able to explore stored data and produce valuable knowledge. Such databases provide a platform for organisations to study past trends in preparation for the future. Decision makers need to be aware of the need to acquire databases in their organisations[207].

[205] (Inmon 2000)

[206] (PhridviRaj and GuruRaob 2014)

[207] (Adriaans 1996)

Chapter 4 *Machine Learning*

As discussed in Chapter 3, the escalation in data generation poses a challenge to organisations in that it takes time to analyse and make sense of the data that they are generating on a daily basis. This calls for advanced data mining tools to help understand the valuable knowledge that comes from databases. Without such Knowledge Discovery Systems, a great deal of valuable information lies unexplored in the databases of various organisations[208]. Therefore, there is a need to continuously develop new Knowledge Discovery Systems whose main focus would be to explore and uncover useful, but hidden information in the databases[209].

Through the application of these various data mining tools, organisations are able to take advantage of the knowledge conveyed by the discovered pattern to solidify their business strategies, knowledge bases and various projects related to scientific research. It is important to note that, without proper data mining tools, the ever-increasing space between data generation and our ability to understand it will forever deny us an opportunity to change the 'data tombs' into 'golden nuggets[210]'.

It should also be noted that data mining tools make use of the machine learning algorithms which are responsible for performing specialised activities. These machine learning algorithms have been tailored to perform a specific type of problem solving. It would therefore not only be senseless but a financial disaster for organisations to apply wrong machine learning algorithms without first conducting a proper investigation[211].

It is therefore advisable that when planning to conduct the application of DM, using various tools, to be aware that individual machine learning algorithms display unique qualities that are suited to solving particular problems. Furthermore, this means that there is a strong relationship between the chosen machine learning algorithm and the problem to be solved[212].

Machine learning algorithms fall into one of two categories, namely supervised and unsupervised learning as discussed next. To be specific, the main emphasis is on classification and clustering learning algorithms which are the main role players in this study.

---

[208] (Mehta and Dang 2011)

[209] (Khandar and Dani 2010)

[210] (Han and Kamber 2001)

[211] (Sung, Chang and Lee 1999)

[212] (Adriaans 1996)

A distinction between supervised and unsupervised learning systems assist decision makers to be aware of the capabilities of each learning system, as well as the value and the capability that each has in sustaining their organisations.

4.1. Supervised learning

Supervised learning is about uncovering the existence of a relationship between independent variables or attributes and a target attribute which is sometimes called a dependent variable[213]. The above relationship is established by identifying and using a selected learning algorithm which will identify those values of the target variable which form an association with predictor variables[214].It is also worth noting that a number of supervised learning systems usually rely on a bigger pool of records which are already pre-labelled, and that this helps the selected algorithms to formulate models that could help in solving the identified problem within an organisation[215].

In addition, as shown in Figure 4.1 below, supervised learning is used to estimate an unknown dependency from known input-output data[216]. This means that the selected input datasets are processed by a learning algorithm and whatever output results that come out of that are compared with those from the sample. Such an action allows for the error signals from the sample output to be at a minimum through repeated adjustments of the learning system[217]. Furthermore, it is worth noting that in supervised learning, the results are assessed, using intrinsic ways due to the presence of class labels which are pre-determined[218].

[213] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

[214] (Larose 2005)

[215] (Larose 2005)

[216] (Ahlemeyer-Stubbe and Coleman 2014)

[217] (Ahlemeyer-Stubbe and Coleman 2014)

[218] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

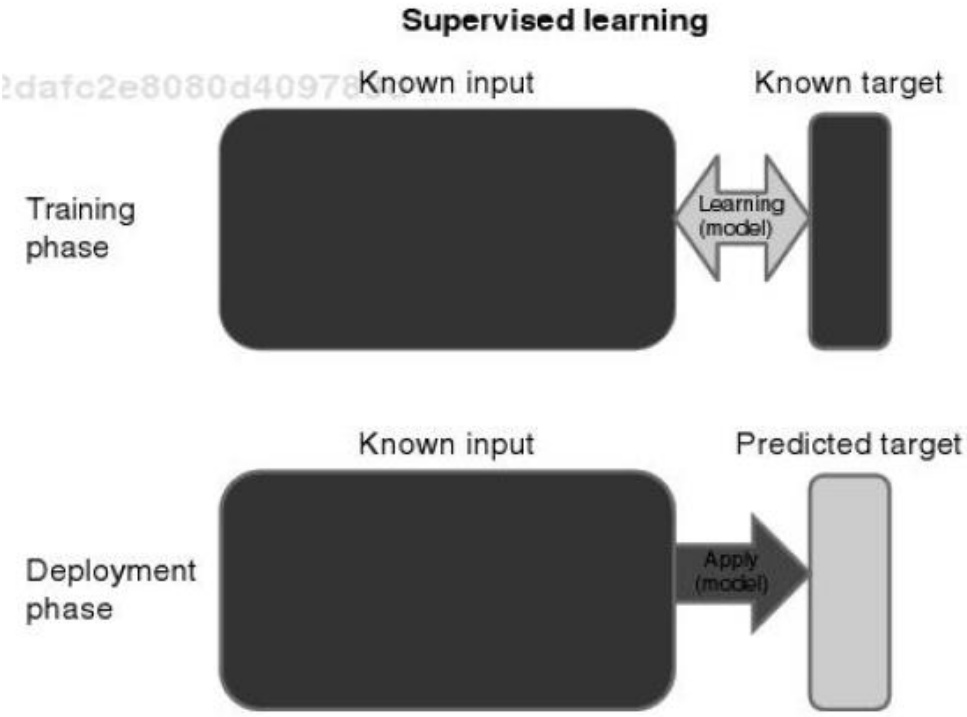


Figure 4. 1 Supervised learning[219].

Supervised learning is popular in many fields of study such as in banking where they may be used to determine whether a housing bond application could be labelled as either bad or good credit risk. In education, for example, students can, through the use of supervised learning, be placed in a special-needs programme based on the displayed patterns of performance or behaviour. Noticeable among the classification methods are the categorical variables which are segmented into pre-determined classes or categories[220]. Due to its success in predicting the value of a target attribute, classification has also been embraced in many areas such as marketing and manufacturing[221].

4.1.1 Classification learning algorithm: an example of supervised learning

A classification machine learning algorithm and, for the purposes of this study, the C4.5 (j48), is the most-used supervised learning due to its practical nature and its ability to influence a number of areas

[219] (Larose 2005)

[220] (Larose 2005)

[221] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

within the KD field of study[222]. It plays a specific role, namely to generate models which come from pre-determined rules. The pre-determined catergories on which classification methods work are, for example, already labled as either yes or no[223].

Classification, which is popular in many fields of study, ranging from scientific discoveries to financial engineering, is used by many experts to predict a number of areas of interest. Its popularity may be tied to the fact that the developed rules called IF…THEN make it easier for experts to obtain the desired results from the selected datasets in a seamless manner[224]. In a banking environment, for example, customers are granted or denied credit based on their risk profiles which may be classified as either low or high[225]. The end result during the use of classification algorithms on any selected dataset are models which are normally evaluated in order to give credibility to the selected data[226].

The whole process of classification, as it is the case with the use of C4.5 (j48) in this study, is preceded by the division of the chosen data into the training and test sets. According to Hamalainen and Vinni,

> A training set is given to a learning algorithm, which derives a classifier. Then the classifier is tested with the test set, where all class values are hidden. If the classifier classifies most cases correctly, it can be assumed that it will also work accurately on future data. On the other hand if a classifier makes too many errors we can assume that it is a wrong model[227].

In an educational environment, classification is popular because of its iterative nature which means that data can be manipulated several times until the desired model is generated[228]. It also has the ability to produce models which, in relation to students' performance, may be categorised as either bad or excellent. In addition, predicting the possibility of students choosing a particular course depends on certain criteria which, when put to the test, would further predict the type of decision that the student will make[229]. Furthermore, classification learning algorithms may be used by teachers in order to group students according to their knowledge, motivation, and behaviour. They can also use

[222] (Maimon and Rokach, Data mining and knowledge discovery handbook 2005)

[223] (Luan 2002)

[224] (H. Du 2010)

[225] (Refaat , Data Preparation for Data Mining Using SAS. 2007)

[226] (Sankar 2004)

[227] (Hamalainen and Vinni 2011)

[228] (Hamalainen and Vinni 2011)

[229] (Ranjan and Khalil 2008)

them during examinations where exam answers are assessed by using specific evaluation criteria on individual marks[230].

Finally, in an educational environment, automatic classification becomes an unavoidable option during the application of the intelligent tutoring systems and adaptive learning environments. It is, therefore, common practice for the selected system to classify learners based on their current situation. This leads to the generation of a model which has the ability to predict the class value from other explanatory attributes. Although the production of such models in small-scale classroom teaching is possible, the computerised learning systems can cover a wider area and collect more data for the chosen classifiers[231].

4.2 Unsupervised learning

Contrary to supervised learning, unsupervised learning systems can search independently for new or previously-unknown knowledge patterns. They are also known for not relying on any target or variables. The selected input data goes through a learning system without any validation against any output. Unsupervised learning systems have the ability to uncover 'natural' structures in the input data. Furthermore, unsupervised learning independently discovers and presents output results whose assessment is often regarded as intrinsic[232].

In addition, as shown in Figure 4.2 below, unsupervised learning can be described as a bottom-up approach where data is allowed to discover independently without pre-determined rules[233].Unlike supervised learning, unsupervised methods do not have any targeted output, and the identified data is searched with the ultimate aim of discovering patterns[234].

[230] (Hamalainen and Vinni 2011)

[231] (Hamalainen and Vinni 2011)

[232] (Maimon and Rokach, Data Mining and Knowledge Discovery Handbook. 2010)

[233] (Luan 2002)

[234] (Larose 2005)

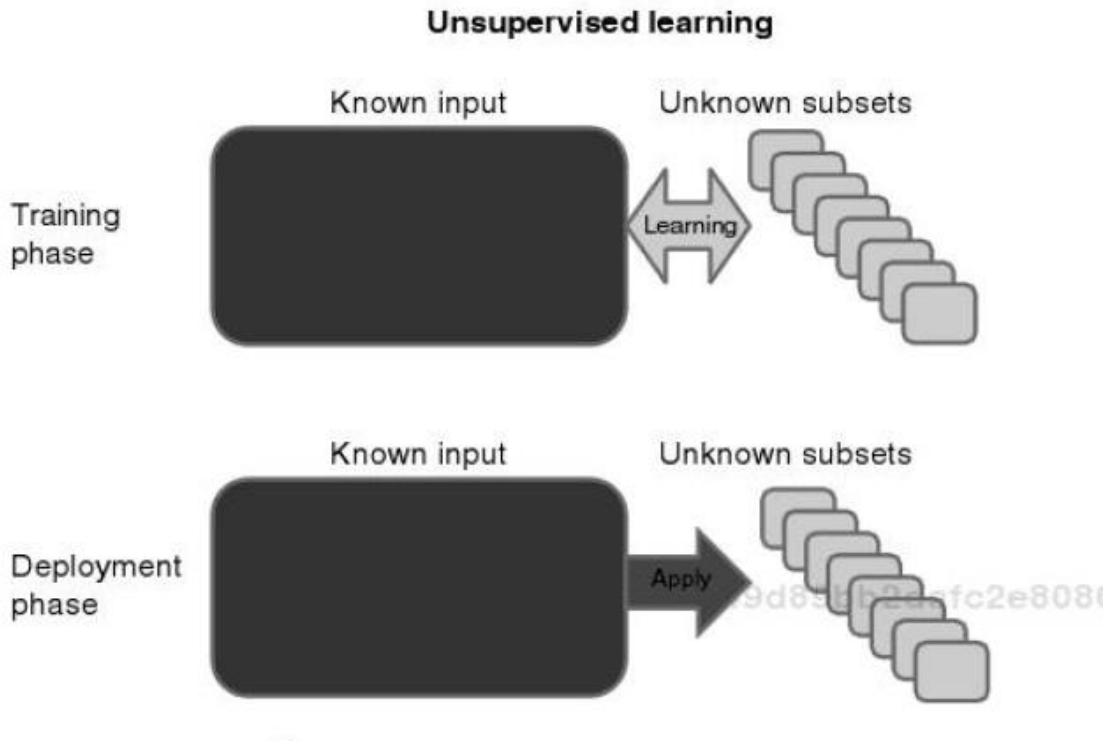


Figure 4.2. 1 Unsupervised learning[235].

Unsupervised leaning, of which clustering learning algorithm forms a part, can be divided into three classes, namely metric distance-based methods, model-based methods and partition-based methods[236].They can also be used as precursors for other DM activities. For example, in neural networks, it is always a good thing when dealing with large sets of data to do clustering first in order to eliminate problems further down the line[237].

In conclusion, unsupervised learning, of which clustering forms a part, is descriptive in nature whereas classification, a member of supervised learning, is predictive in nature. Although there is a thin line between predictive and descriptive learning systems, it is common for the models that come out of these learning systems to display each other's characteristics. Even though there are various reasons for choosing suitable learning systems, KDD tends to favour description rather than prescription[238].

[235] (Larose 2005)

[236] (Fayyad and Stolorz, Data mining and KDD: Promise and challenges 1997)

[237] (Larose 2005)

[238] (U. M.-S. Fayyad 1996)

4.2.1 Clustering learning algorithm: an example of unsupervised learning

When it comes to data analysis and interpretation, the clustering algorithm emerges as one of the most widely-used research tools[239]. What makes clustering distinct is its ability to uncover hidden data without any supervision[240].

A clustering learning algorithm's main goal (and for this study, the SimpleKMeans) is to partition groups of items based on similarities and, in turn, gain a deeper understanding with regard the causes of such similarities and differences between groups. The word 'similar' is the most accommodating term in the sense that it is impossible to provide a detailed explanation regarding the characteristics of given attributes. On the other hand, it is regarded as a difficult exercise to look for accurate comparisons between clusters and, therefore, the results are usually subjective in nature[241]. Due to difficulties when it comes to giving an accurate analysis of the clusters, clustering relies on heuristics which give near-perfect results[242].

In addition, the clustering learning algorithm can also be referred to as the act of 'finding islands of simplicity in the data[243]'. Furthermore, the clusters have noticeable features which include homogeneous characteristics which result in data objects within a cluster being referred to as members[244].

Based on the above explanation, clustering can be accomplished by following 'the principle of maximizing the intra-class similarity and minimizing the interclass similarity'. This means that clusters are caused by data objects with observable similarities which are usually non-existent when compared with data objects which have been formed in other clusters. It is, therefore, through clustering that a taxonomy can be developed where classes bearing similarities are grouped together[245].

As mentioned above, the use of subjectivity, especially in fields such as biology, is common where models are generated by grouping together those subsets of data that display common characteristics such as structure, chemical composition and functions. The numbers of clusters which are formed

[239] (Cios 2007)

[240] (Cios 2007)

[241] (Michaud 1997)

[242] (Michaud 1997)

[243] (Refaat, Data Preparation for Data Mining using SAS. 2007)

[244] (H. Du 2010)

[245] (Han and Kamber 2001)

through clustering process are normally unknown as, by nature, clustering is an unsupervised approach which is expected to produce unknown but useful results[246].

As previously mentioned, a clustering learning algorithm such as the SimpleKMeans helps to partition and group together instances that show similar traits. EDM, for example, may apply clustering techniques to divide learners into specified groups based on their learning or cognitive patterns[247].Through clustering, an education institution may also develop homogeneous models based on elements such as age, background and areas of interest[248].

Furthermore, one of the areas in which clustering is popular is on time series which involves putting together all those data objects that bear similar features. A good example is in the financial sector where stocks that continue to display similarities over time are grouped together[249]. Finally, this age-old human activity called clustering has been widely applied in many areas such as pattern recognition analysis of data, image processing and in market research. It is, therefore, through clustering that data distribution and segmentation into either densely- or sparsely-populated areas can be identified as observed over a period[250]. The application of both supervised and unsupervised learning systems may help organisations to obtain a detailed description and to predict the future of the data trends[251]

4.3 Conclusion

Using machine learning algorithms has become one of the critical missions of enterprises. Since not every user completely understands the theory of data mining, choosing the best machine learning solution is not easy[252]. The availability of different DM tools which make use of supervised and unsupervised learning continues to offer a much-needed solution to understanding the ever-increasing data in our universe. Their uniqueness and flexibility, together with ease of understanding, allows people from different fields of study to apply them in their quest for knowledge. Such tools further offer users a variety of learning algorithms which are uniquely designed to perform specific functions

[246] (Dibl and Carbonel 2012)

[247] (Bousbia and Belamri 2014)

[248] (Ranjan and Khalil 2008)

[249] (Mehta and Dang 2011)

[250] (Han and Kamber 2001)

[251] (Han and Kamber 2001)

[252] (Seng and Chen 2010)

which may, ultimately, help organisations to perform better in the competitive environment in which they are operating.

Chapter 5 *Free State Department of Education's data sources*

With reference to section 1.1 (Chapter 1), the process of conducting Grade 12 examinations, and the subsequent collection, storage and distribution, is governed by legislation which is duplicated across the nine provincial departments in the country. The FSDoE is, therefore, expected to provide the infrastructure to establish data sources which would not only meet the specified requirements but also provide a platform for members of the public to access such data whenever a need arises.

According to Gxwati,

> "the Free State Department of Education is expected to report and account to parliament and to the public on whether the resources allocated to the department are being utilised in an efficient and cost effective manner as guided by constitutional and legislative frameworks[253]".

This means that through the FSDoE, the provincial government has the task of ensuring that the Grade 12 data results are not only credible but also kept in a safe and easy-to-access database.

The EMIS database, from which the copy of the data that is used in this study comes, is managed by the section within the FSDoE that goes by the same name: EMIS. In this study, the EMIS database and EMIS section are, interchangeable, refer to the database and section managing the database, respectively. In preparation for this study, and after making a formal request, EMIS section made available the 2010 to 2013 FSDoE's Grade 12 results data.

5.1 The nature of the selected data

In consideration of the ethical aspects, the copy neither includes the identity numbers nor the names of the candidates who wrote the Grade 12 examinations during the years selected for use in this study.

The copy provided is in an excel format and covers Grade 12 data results for students who wrote exams during the years 2010 to 2013, and is divided into learner levels per school, learner levels per subject, and subject averages. All the five districts are included in the copy, but for this study, only Motheo District is quoted. The selected data had the following original attributes:

- District code: DC 17 is similar for all Motheo District schools;
- District: Motheo;

[253] (Gxwati 2011)

- EMIS Number: differs according to schools;
- School name: these are in alphabetical order;
- Subject code: unique number which is similar to all schools doing the subject;
- Subject name: in alphabetical order and common in schools doing it;
- Data year: 2010 to 2013;
- Learners 0-29%: number of learners within the percentage bracket;
- Learners 30-39%: number of learners within the percentage bracket;
- Learners 40-49%: number of learners within the percentage bracket;
- Learners 50-59%: number of learners within the percentage bracket;
- Learners 60-69%: number of learners within the percentage bracket;
- Learners 70-79%: number of learners within the percentage bracket; and
- Learners 80-100%: number of learners within the percentage bracket.

The data provided consists of subjects written in a specific year and as observed over a four-year period. For example, in Table 5.1 below, Academy of Excellence, an independent combined school in the Motheo District has the following data for Afrikaans (First Additional Language):

Table 5. 1 Performance of learners across percentage brackets.

Data year	Learner 0-29%	Learner 30-39%	Learner 40-49%	Learner 50-59%	Learner 60-69%	Learner 70-79%	Learner 80-100%
2010	2	8	8	4	0	0	0
2011	0	8	7	7	2	1	0
2012	2	9	5	4	3	0	0
2013	0	3	8	2	2	0	0

From the above table, for example, in 2010 the Academy of Excellence CI/S had a total of 22 learners who wrote Afrikaans (First Additional Language). Of the 22 learners, two fall into the 0-29% bracket, eight into the 30-39% bracket, eight into the 40-49% bracket, four into the 50-59% bracket, and none in the upper three brackets, namely 60-69% , 70-79% and 80-100%.

In the Motheo District, there are 88 schools that offered Grade 12 during the selected period of this study. They are categorised as public, independent or specialised schools. They are located in the city of Bloemfontein, towns surrounding Bloemfontein and in the townships. They form a district called Motheo. The subjects that have been selected for this study are English Home Language (EHL), csv,

Mathematics and Mathematical Literacy, and are mostly offered unevenly across the schools in the Motheo District. This means that some schools offer all subjects while others are selective, for example:

- Albert Moroka High School offers only English First Additional Language, Mathematical Literacy, Mathematics, and Mathematics: probability; data handling
- Brebner Secondary School offers English Home Language, Mathematical Literacy, Mathematics, and Mathematics: probability; data handling

The selected subjects across the schools in the Motheo District have the following data which represent the number of learners who wrote an exam in a particular subject:

Table 5. 2 English First Additional Language 2010-2013.

Name of school	2010	2011	2012	2013
Albert Moroka S/S	92	86	85	65
Atlehang S/S	135	159	148	135
Bainsvlei C/S	30	24	30	27
Bartimea Special	13	11	13	11
Bloemfontein S/S	92	79	93	67
C & N H Meisieskool Oranje	127	122	93	134
Christiaan De Wet C/S	39	38	45	48
Christiaan Liphoko I/S	63	34	41	52
Commtech C/S	115	161	130	128
Dr Viljoen C/S	52	45	35	69
Dr Blok S/S	118	90	74	94
Excelsior C/S	22	31	19	18

In Table 5.2 above, it can be seen that 92, 86, 85 and 65 learners from Albert Moroka S/S wrote the English First Additional Language Examinations in 2010, 2011, 2012 and 2013, respectively. Furthermore, Table 5.3 below shows that 22, 25, 22 and 15 learners from the Academy of Excellence wrote the Grade 12 English Home Language examinations in 2010, 2011, 2012 and 2013,

respectively. In both Table 5.1 above and Table 5.3 below, Bloemfontein S/S offered both subjects at Grade 12 level during the period under investigation.

Table 5. 3 English Home Language 2010-2013.

Name of school	2010	2011	2012	2013
Academy of excellence	22	25	22	15
Accelerated Christian college	53	55	31	50
Bloemfontein S/S	46	34	42	58
Bloemfontein South High	100	85	86	81
Brebner S/S	194	176	173	198
Calculus Bloemfontein C/S	56		47	71
Castlebridge CI/S	61	64	33	42
Christian Brothers College	29	26	24	29
Eunice S/S	164	155	162	185
Grey Kollege	77	78	68	73
Headstart High	21	40	31	21
HTS Louis Botha	181	176	208	246

The data shown in Tables 5.2 and 5.3 above can be broken down further according to the levels of achievement in a subject per school as shown in Table 1. In order to collect, store and access the data above, a number of stages have to be followed.

5.2 Grade 12 examination processes in the FSDoE

With reference to section 1.1 (Chapter 1), which states that annually, the FSDoE conducts an examination process for Grade 12 learners, it is therefore required that the FSDoE submit a management plan which covers the entire examination process. The management plan commences with the appointment of examiners and internal moderators who are responsible for setting and

moderating the question papers, and concludes with the release of the results[254]. This process starts 24 months prior to the commencement of the actual examinations, and ends six months after the exams have been written. The individual students' exam papers are marked based on criteria set out by the FSDoE.

5.2.1 Learner achievement levels

This study makes use of the Grade 12 results which are described, using the seven identified levels of performance per subject. These levels summarised in Table 5.4 below, guide teachers in assessing learners and assigning the correct rating. The consolidated mark for each learner determines the type of certification a learner is to be awarded by Umalusi.

Table 5. 4 Scale of achievement for the National Curriculum Statement Grades 10-12 (General).

Rating code	Achievement description	Marks %
7	Outstanding achievement	80-100
6	Meritorious achievement	70-79
5	Substantial achievement	60-69
4	Adequate achievement	50-59
3	Moderate achievement	40-49
2	Elementary achievement	30-39
1	Not achieved	0-29

5.2.2 Certification

After collecting and verifying the individual learners' marks, the FSDoE's examinations and assessment section has the responsibility of handing over all 'the certification records to the historical certification records section of the Department of Education'. The EMIS section of the FSDoE has been assigned the task of ensuring that the storage of data records such as the Grade 12 results happens within three months after the certification of records has been approved, and it is only under exceptional circumstances that an extension may be granted.

[254] (South Africa. Department of Basic Education 2005)

5.3. EMIS in the Free State Department of Education

In addition to the activity mentioned in 5.1.2, the FSDoE is expected to secure a place for 'back-up copies of all the historical certification records of provincial assessments including the Grade 12 learners' marks which are part of this study[255]'.

As with all the provincial EMIS sections, EMIS in the FSDoE must ensure that accessing these historical records becomes an integral part of, and cannot be divorced from, all the activities of the examinations and assessment directorate. It is therefore the responsibility of the EMIS section in the FSDoE to ensure that a proper computer infrastructure is in place which will allow access to the centralised historical database. By having such an infrastructure in place, EMIS in the FSDoE offers an important service to both internal and outside stakeholders which allows for the records to be accessed and used for queries, the combination of results, and for the verification of certification data[256].

Although EMIS deals with all the data produced within and outside the FSDoE, some parts of the data are not made public and, like the data used in this study, namely Grade 12 data, permission is granted upon receipt of a formal request.

The EMIS section in the FSDoE makes it easier for individuals to access data on schools in the province which includes schools, teachers and learners. This allows the FSDoE decision makers easy access to updated data on a regular basis to ensure proper planning, improvement in service delivery and accurate projections for the future[257].

Due to the sensitive nature of the data that is stored in the database, the responsibility rests with the section responsible for managing the EMIS database in the FSDoE to not only comply with the requirements set by Umalusi but also to keep a constant eye on the database by monitoring and managing the computer system which must conform to the minimum requirements for a computer program.

The technology that forms the EMIS framework is listed as:

- SQL server 2005 database;
- a dashboard manager for performance indicators;
- SQL replication which helps to push the data to the web service;

[255] (South Africa. Department of Basic Education 2005)

[256] (South Africa. Department of Basic Education 2005)

[257] (Kok n.d.)

- GIS: ARC IMS viewer; and
- Web services which run on Apache Tomcat[258].

The EMIS section has an additional responsibility of ensuring that the collected data is stored in an EMIS database with a standardised format which is approved by both the quality assurance body, Umalusi, and nationally by the Department of Basic Education[259]. The EMIS database allows access to detailed data on the level of a specific school, in a specific district in the province, for a particular period.

5.3.1 Coding standards

As indicated in section 3.2.1 (Chapter 3), relational databases follow specific coding standards which not only continue to make them popular, but have stood the test of time. The EMIS database is also designed according to specified data coding standards which stipulate that as a relational database, it must have all the tables linking via a primary foreign key relationship. In addition, it is premised on the principle which requires that the links may be 'one-one', 'one-many', 'many-one' or even 'many-many[260]'. In Figure 5.3.1 below, it can be seen that the entity, Grade 12 students' marks, as it is the case with EMIS in the FSDoE, has a relationship with three attributes, namely examination number, subject and learner achievement. In the EMIS database, an individual learner is identified by, among other attributes, a unique examination number, subjects enrolled for and achievement levels per subject. The cardinality between a particular student and the selected attributes, in this case, is one-to-many.

[258] (Kok n.d.)

[259] (South Africa. Department of Basic Education 2005)

[260] (South Africa. Department of Basic Education 2005)

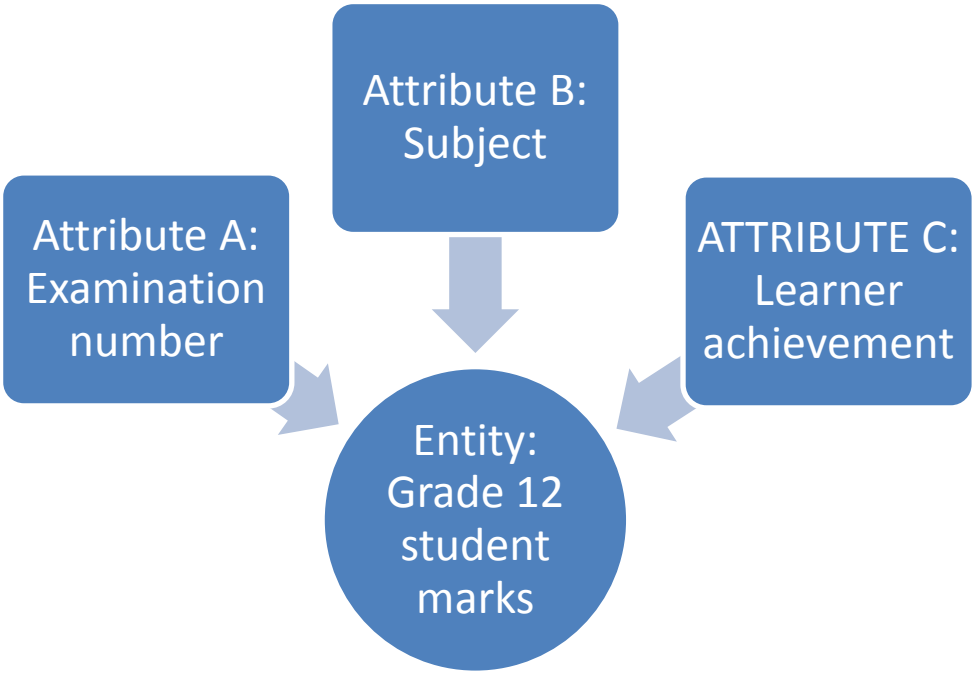


Figure 5.3. 3 The one-to-many relationship in the EMIS database.

In addition, based on the relationship which exists between entities and attributes, the EMIS database also follows the principle of using primary and foreign keys in order to avoid confusion during the design and subsequent search for data by the user. For example, in Figure 5.3.2 below, the EMIS database recognises the examination number as a unique number which cannot be duplicated. Furthermore, a student whose examination number is 1, as shown in Figure 5.3.2 below, forms a primary key which is used by the other attributes as a point of reference. For example, a student with the examination number 1 who has written Mathematics and English, may have taken Maths Literacy, pure Mathematics and English Home Language. Another student in the same school may have taken the same combination but the difference will be on the primary key which uniquely identifies the students as two, unique people.

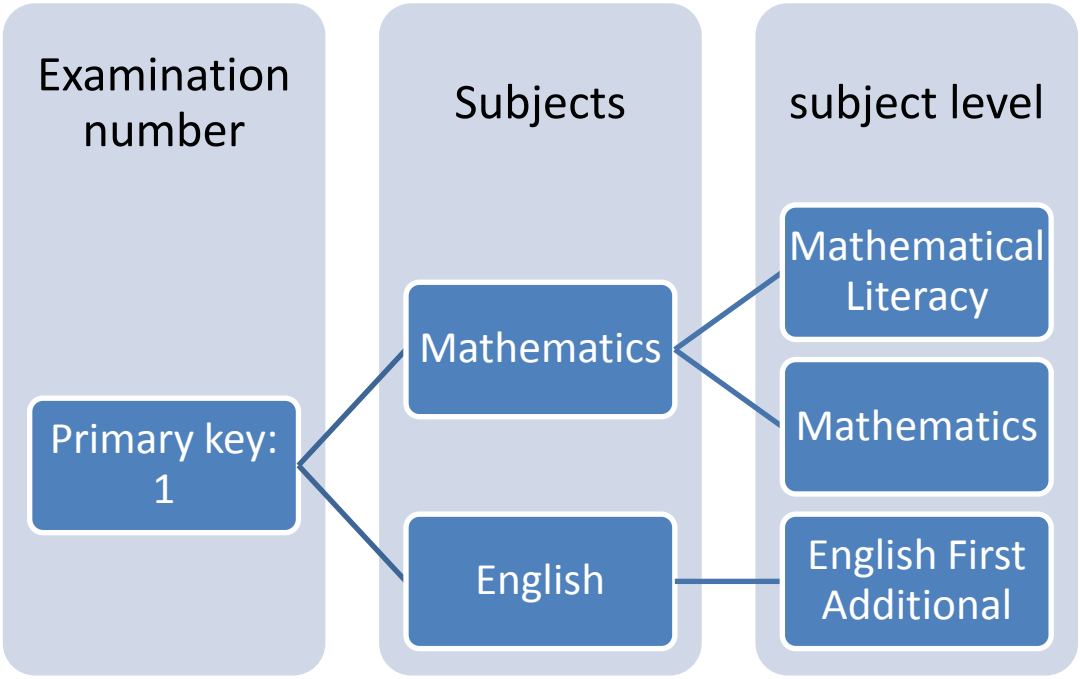


Figure 5.3. 4 The one-to- many relationship in the EMIS database.

EMIS further recognises that such standardisation, which is common in relational databases, is an acknowledgement that quality of data is key and that data has many faces. This means that it has to be relevant to its users. Finally, as is the case with the EMIS copy used in this study, data must be made accessible and easy to interpret[261].

5.4 National Census Data

Census data is a benchmark of demographic information at all levels of geography, on any given place and time for a nation. The results of the 2011 national census provide interesting insights which have a direct influence on the data that the education sector produces on an annual basis. The following were used for this study[262].

Population size

Age-sex distribution

To gain a better understanding of the demographics and socio-economic status of the country, an understanding of the age-sex distribution is crucial. Of the total population, there are more females than males, nationally.

Race distribution

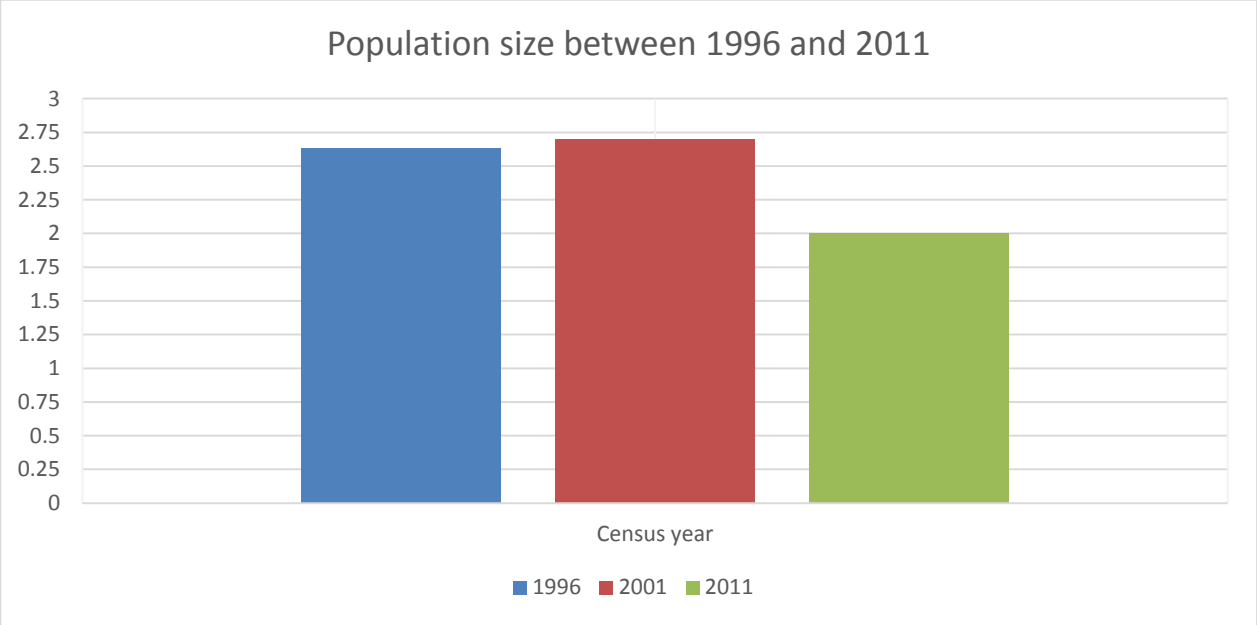
[261] (South Africa. Department of Education 2005)

[262] (Statistics South Africa. 2012)

The black African population has constantly, over the last three censuses, maintained an average of more than 75%.

Migration patterns

The current results from the 2011 census, and as shown in Graph 5.1 below, the Free State (Like the Eastern Cape, Northern Cape, and KwaZulu-Natal) experienced a net out-flow of people as observed over a ten year period. On the other hand, the Western Cape and Gauteng experienced the highest in-flow over the same period.



Graph 5. 2 Population size in the Free State from censuses 1996, 2001 and 2011.

Schooling

The 2011 census indicates that a vast majority of students in South Africa attend public educational institutions. Only 5% attend private institutions. The Free State (6.4%), Western Cape (7.5%) and Gauteng (16.7%) showed the highest increase in private school attendance.

Another notable feature is that 73.9% of the black population attends school in comparison with the 71.8% Indians/Asians and 77.7% whites.

Annual household income

As shown in Table 5.5 below, the 2011 census indicates that over the past ten years, the average annual household income has more than doubled in the country.

Table 5. 5 Average household income amongst racial groups from census 2011.

Race	Average annual income	Average household increase %
Black	R 60 613,00	169.1%
White	R 365 134,00	88.4%
Indian	R 215 541,00	145.2%
Coloured	R 112 17,00	R 118.1%

Table 5.6 below, further indicate that Limpopo had the lowest average annual household income of R56 844, followed by the Eastern Cape where the average was R64 539. On the other hand, the Gauteng province had the highest average annual household income at R156 243, followed by the Western Cape with a figure of R143 460.

Table 5. 6 Average annual income per province.

Province	Average annual household income	Average annual Household %
Limpopo	R56 844	147.3%
Eastern Cape	R64 539	120%
Gauteng	R156 243	98.9%
Western Cape	R143 460	83.6%
Mpumalanga	R 77 609	148.9%
Free State	R 75 312	145.1%
Eastern Cape	R 64 539	120.0%

Housing

As a basic human right, housing directly impacts other important elements such as health, welfare and social status in communities. When it comes to people living in rent-free or fully-paid houses, Table 5.7 indicates that the Free State has 51.6% of such households.

Table 5. 7 Average rent-free housing per province.

Province	Average annual Household %
Limpopo	52.7%
Eastern Cape	51.7%
Gauteng	27.9%
Western Cape	35.3%
Mpumalanga	52%
Free State	51.6%
Eastern Cape	50.7%
Northern Cape	20.9%
North West	20%

Provision of services

The combination of piped water inside the dwelling and outside the yard is high in all provinces except in the Eastern Cape and Limpopo which have the lowest proportions. Dwellings with piped water are highest in the Gauteng province (89.4%), followed closely by the Free State with 89.1 %. The provision of electricity showed a marked increase from 58.2% in 1996 to 84.7% in 2011.

Availability of household gadgets or technologies

Ownership of mobile phones by households in South Africa increased from 31.9% in 2001 to 88.9% in 2011. Households that own computers increased from 85% in 2001 to 21.4% in 2011.

The FSDoE's examinations and assessment directorate has the important task of ensuring that each examination process is followed according to the management plan. The quality assurer, Umalusi, and the affected bodies ensure that all the activities leading to the handing over of the Grade 12 data to the DBE is done within a specified period. Proper infrastructure is the key when it comes to the collection, storage and accessing of Grade 12 results data. The presence of the EMIS database ensures that even beyond the examination processes, such data is made available for further scrutiny. In addition to the Grade 12 results data, the presence of census data is able to provide a clearer picture of the state of affairs in the country as observed over a specified period.

Chapter 6 *Research methodology*

As mentioned in section 2.4, this study is guided by the CRISP-DM methodology, an industrial-standard mining methodology which has been discussed extensively.

According to the CRISP-DM methodology, the life cycle of a data mining project consists of six phases:

Business understanding

Data understanding

Data preparation

Modelling

Evaluation and

Deployment[263]

6.1. Business Understanding

The controversy surrounding the credibility of Grade 12 results data mentioned in chapter 1.1 influenced the need to undertake this study on FSDoE' s Grade 12 results data. In addition, the mandate assigned to Umalusi and EMIS, to have clearly drawn education standards which lead to credible and reliable data, need to be put to test in order to allay fears that emanate from some individuals and institutions that cast doubts on the state of Grade 12 results data. It is through this study that an application of DM tools and techniques using internationally recognized software called WEKA was identified. The use of Grade 12 results data from the Free State which spans over a four year period is intentionally chosen to establish if there are any patterns to be identified from such data. Furthermore, it is hoped that credibility and reliability of data results produce positive influence on how top managers, and in this case, FSDoE, make informed decisions based on the models generated during the application of selected DM tools and techniques on the Grade 12 results data. Finally, through this study, such decisions are not only be based on the tail-end of the education processes but be tied to the relevant education theories that would have a positive influence, and thus leading to a thriving economy and progressive citizenry. Data understanding, as it the case with this study, is

[263] (H. Du 2010)

directly tied to the state of affairs in the country and this is one of the reasons 2011 census data was incorporated to make sense of the models generated using WEKA software. Furthermore the use of theories in education, and as it is the case with this study, Bloom's Mastery Learning model and Leaning organizations, are all forming an integral part aimed at making sense of the FSDoE'S Grade 12 results.

6.2 Data understanding

As mentioned in section 2.5, this study makes use of the 2010-2013 Grade 12 results data obtained from FSDoE's EMIS database which is managed by a section with the same name EMIS. Even though data from all subjects was provided, only English First Additional Language, English Home, Mathematical literacy and Mathematics subjects were selected to be part of this study. The main focus is to ascertain if there are any patterns to be identified, from using WEKA as a DM tool, on the achievement levels of learners who set for Grade 12 examinations in the Free State during the specified period of time. A detailed account of the selected data is given below.

6.2.1 Collecting and integrating data

The selected data in this study comes from two sources which are EMIS database from FSDoE and from 2011 National Census. The former was used during the application of DM tools and techniques using WEKA software whilst the latter was used to contextualize and make sense of the models that were generated. This further means, census 2011 gives a national perspective whose influence cannot be divorced from the models generated during the application of DM tools and techniques on the Grade 12 results data. The data represents marks obtained by learners who sat for Grade 12 examinations in the Free State from 2010 up to 2013. For the purpose of this study, only data dealing with Free State province was selected from the current 2011 census data which was obtained from Statistics South Africa.

6.2.2 Data Description

With reference to section 2.5, and it has been confirmed from FSDoE's database, the data contains the following fields:

District code: a code for schools in a particular district (nominal and numeric)

District: name of the district under which the schools fall (nominal)

EMIS number: unique school identifier (numeric)

School: name of school (nominal)

Subject code: a unique identification number (numeric)

Subject name: name of a particular subject (nominal)

Data year: the year the data was produced (numeric)

Learners within a percentage bracket: number of learners in a specified percentage (Numeric)

6.3 Data preparation

As it was mentioned in chapter 2.4.3, more time was allocated to the preparation of the selected FSDoE's Grade 12 data results. In line with the requirements of the CRISP-DM methodology, the following sub-steps form part of the entire data preparation which was done on the selected Grade 12 results data.

6.3.1 Data selection

From the original data obtained from the FSDoE database, the following attributes were available: district code, district name, EMIS no, school name, subject code, subject name; data year and learners within a specified percentage bracket. Out of the eight original attributes, four have been removed as they served no purpose in the study. These attributes are district code, district name, EMIS no and subject code. The three included attributes are location, performance category and overall school performance.

The FSDoE's website has a webpage called the school finder where all the schools in the province are listed including their location. This helped in identifying the exact location of the school as their position plays an important role in this study.

The candidates' names, surnames, identity numbers attributes were removed prior to receiving the copy of the database to protect the privacy of the learners who wrote Grade 12 from 2010 up to 2013 in the FSDoE. The district code and name are similar for all schools and play no part in the study and therefore were also removed. The EMIS number and subject code do not have an impact on the study and have been removed. The inclusion of the location in the study is to highlight the position of the school in the province as city, town or township school.

Table 6. 6 overall performance of candidates in the Free State 2010-2013 Grade 12.

Examination Year	Total who wrote	Total achieved	% achieved
2010	27 586	19 499	70.7
2011	25 932	19 618	75.7
2012	24 265	19 676	81.1
2013	27 105	23 689	87.4

Looking at table 6.1 above: in 2010, the Free State Province's pass rate was 70.7% and it is this pass rate is used to measure individual school's achievemet during this specified examinations period.

Table 6. 7 comparison of schools' performance in the Motheo district in 2010 to the province's average performance.

School Name	Total who wrote	Total achieved	% achieved
COMMTECH Commercial Secondary School	114	71	62.3
Calculus Bloemfontein SI/School	49	42	85.7
Fichardt Park Secondary School.	204	203	99.5

If one looks at table 6.2 above, in 2010, COMMTECH Commercial secondary school's pass rate was 62.3% and that is Below the Average performance in the Free State which is 70.7%. On the other hand, Calculus Bloemfontein SI/School and Fichardt Park Secondary School,

obtained 85.7% and 99.5%, respectively. This is Above the Free State province's 70.7% Average performance.

Table 6. 8 Achievement levels allocated to Grade 12 subjects data results.

Achievement levels	Achievement levels %
A	80-100%
B	70-79%
C	60-69%
D	50-59%
E	40-49%
F	30-39%
G	0-29%

As indicated in Table 6.3 above, learners'performance is allocated across seven levels of achievement with 0-29% being the lowest and 80-100% being the higest. The same achievement levels can be interpreted using letters with G representing the lowest and A representing the highest level of achievement.

In preparation for the application on WEKA, of the selected FSDoE Grade 12 data, the performance categories are divided according to the following percentages:

- Below 40
- Below 50
- Below 60
- Below 70
- Below 80

The raw data for 2010-2013 Grade 12 Motheo district Grade 12 schools results have been converted from original number of learners per percentage category to percentage of performance per category. In order to understand the categorization mentioned above, the following example is from one of the selected schools.

In 2010, the data from Albert Moroka High school learners' performamnce in Mathematics results were:

Level G = 26%

Level F = 41%

Level E = 19%

Level D = 5%

Level C = 5%

Level B = 3%

Level A = 2%

The above, and as indicated in table 6.3 earlier, indicates that 26% of the learners were in the achievement level G (0-29%) category whereas only 2% (80-100%) of the learners were in the achievement level A.

Instead of looking at percentage in each level, the new groups are as follows:

Below 40% = 26 + 41 = 67%

Below 50% = 21 + 41 + 19 = 86%

Below 60% = 21 + 41 + 19 + 5 = 91%

Below 70% = 21 + 41 + 19 + 5 + 3 = 94%

Below 80 % = 21 + 41 + 19 + 5 + 3 = 2 = 96%

With reference to the five created performance categories (from Below 40 to Below 80) and the 2010 data which has been used as an example, 67% of the learners at Albert Moroka High School had less than 50% whilst, 86% of the learners had less than 60%. This combination allows us to compare the schools' pass % to that of the province for a particular year to build our performance classification attribute.

Furthermore, the schools' overall pass rate is compared to the Free State province's Average pass rate. Based on the above, the schools' overall pass rate was classified as Below, Average or Above.

For classification, the chosen classifier algorithm is C4.5 and it is implemented using the classifier class j48. In addition, each selected subject's dataset is divided into training and test

options. In preparation for clustering using SimpleKMeans algorithm, the same datasets for each subject are subjected to clustering algorithm using 5 clusters respectively.

The subjects were intentionally selected as they are offered in most schools throughout the province. They were selected to see if there are any pattern to be drawn by comparing the Grade 12 leaners' performance in the same subjects during the 2010-2013 examinations periods.

6.3.2 Data cleaning

The FSDoE examinations and assessment database which is part of this study contains secondary data that has been provided by the directorate responsible for the entering, release and storage of the Grade 12 results. The cleaning and removal of incomplete data was done on those schools whose results data appears as duplicates for the same subject and year. For example 2010 Mathematical literacy dataset for one school is repeated in the subsequent row and it contain the same data as that which appears in the previous row.

6.3.3 Data formatting

In order to meet the WEKA's data formatting requirements, the selected data for this study has been prepared on a spreadsheet. The data which is in a spreadsheet has been saved using a comma separated value (csv) format.

Although WEKA's data storage method is ARFF format, it is easy to convert data from a spreadsheet to ARFF. In addition WEKA can read csv spreadsheet directly, thus saves time as there is no need to create the ARFF file. All the selected data from the FSDoE examinations and assessment database has been saved using the csv format. All the selected subjects' dataset (English Home Language, English First Additional Language, Mathematics and Mathematical Literacy,) from all the selected years (2010 to 2013) have been separately saved in a csv format which complies with WEKA's requirements.

6.4 Data modelling

For this study, 3.8.0 version of Waikato Environment for Knowledge Analysis (WEKA) has been selected to generate models using FSDoE's Grade 12 results data. Tied to the choice of WEKA further meant an identification of one supervised and one unsupervised method for use in this study. For supervised method, a C4.5 decision tree classifier using j48 algorithm and for unsupervised methods SimpleKMeans clustering formed part of this study. The motivation for

choosing C4.5 decision tree learner using j48 algorithm is discussed earlier in chapter 4.1 whilst that of choosing, SimpleKMeans clustering was discussed in section 4.2.

The following four datasets were generated:

- English Home Language (GR12 2010-2013)
- English First Additional Language (GR12 2010-2013)
- Mathematics (GR12 2010-2013)
- Mathematical Literacy (GR12 2010-2013)

As mentioned earlier, the datasets from the above selected subjects are further divided into training and test options using WEKA's unsupervised instance filter called Resample. Parameters were set in order to divide the data sets into two. The percentage for the training data set is 66% whilst the test data sets is 33% as illustrated in Table 6.4.

Table 6. 9 Data set allocation into 66% training and 33% testing: selected subjects

Subject	Data set allocation	Number of instances
Mathematical Literacy	Train	191
	Test	97
	Total	288
Mathematics	Train	181
	Test	91
	Total	272
English First Additional Language	Train	165
	Test	83
	Total	248
English Home Language	Train	45
	Test	23
	Total	68

The training data set per subject is used to develop a model for predicting overall school performance as either Above, Below or Average compared to the provincial average. The test data set used for testing the accuracy of the model generated during the training phase.

For clustering purposes, SimpleKMeans algorithm using k-means algorithm has been selected to be applied on all the selected subjects' data sets. The selected data sets which are in csv format are as follows:

Grade 12 Motheo English Home Language 2010 to 2013

Grade 12 Motheo English First Additional Language 2010 to 2013

Grade 12 Motheo Mathematics 2010 to 2013

Grade 12 Motheo Mathematics Literacy 2010 to 2013

Table 6. 10 Data sets selected for clustering using SimpleKMeans algorithm

Subject	Data set allocation	Number of instances
Mathematical Literacy	Train	288
	Total	288
Mathematics	Train	272
	Total	272
English First Additional Language	Train	248
	Total	248
English Home Language	Train	68
	Total	68

The above datasets for each subject are used to develop 5 cluster models. This means, in total, there are four 5-cluster models generated from the four selected subjects.

6.4.1 Data Mining Tool Selection

With reference to the research question and objectives in chapters 1.21 and 1.2.2 respectively, the goal of data mining task in this study is to establish credibility in the Grade 12 results data. The elimination of controversies around the data generated from Grade 12 results can be stopped by selecting a data mining application software, like WEKA, which is compatible with universally accepted CRISP-MD methodology.

With further reference to chapter 1.3, WEKA is chosen for this study for, amongst many reasons, the installation process which is quite straightforward. Furthermore, WEKA's compatibility with Microsoft Windows has influenced the decision to choose it for this study[264]. In this study of the FSDoE Grade 12 2 results, WEKA is operating on MS Windows operating system[265].

WEKA has a variety of algorithms which can be specifically chosen to solve and suit the needs of any organization and for this study C4.5 and SimpleKMeans algorithm have been selected respectively to answer the research question raised earlier in chapter 1. The choice of these two algorithms, and as explained in chapter 4 offer this study a chance to apply both supervised and unsupervised learning on the selected data. WEKA's ability to operate on csv format made is easy to convert the selected FSDoE's Grade 12 data sets to suit this important requirement. The variety, quality, and flexibility of visualization tools have strongly influenced the usability, interpretability, and attractiveness of a data mining system like WEKA for this study[266].

6.4.2 Data Mining Test Design

In this study and for classification purposes, the datasets from all the selected 2010 to 2013 Grade 12 subjects are divided into training and testing data sets. For clustering, the results are obtained through the application of SimpleKMeans algorithm and using 5 clusters. The use of 2010 to 2013 Grade 12 results allows for the generated models to be compared with each other using similar subjects.

6.4.3 Building and assessing models using data mining

From the classification models that were generated using the classifier J48 on the training data sets, visual representation is used to analyse the models. In addition, the testing data sets, which produced results from the training models, make us of the Confusion Matrix to test the accuracy of the results per subject as observed over a four year period from 2010 to 2013 Grade 12 examinations.

---

[264] (H. Du 2010)

[265] (Chen and Wei 2009)

[266] (Chen and Wei 2009)

In order to prove credibility of the FSDoE'S Grade 12 results' raw data is compared with the class attribute which has bee derived from Umalusi's results. The comparison is of the school's performance (in percentage) to that of the province in a particular year.

The clustering models look at the following:

Number of instances within clusters

Data years within different clusters

The performance categories in percentges

Location of the schools

Overall school performance per category

In this study on the FSDoE's examinations and assessment database, the output emanates from the different model-building activities which are iterative in nature. They are, later going to be analysed individually and presented for further decision making.

6.5 Evaluation Phase

From this study and in reference to section 2.4.5 and the research question in section 1.2, at this stage, the models that were generated after the application of DM tools and techniques on FSDoE's Grade 12 results data are evaluated in terms of achieving the prior stated aims and objectives. Models generated from 2010 to 2013 Grade 12 Grade 12 results data are measured against each other as seen over period of four years. The following, are sub-phases that form part of the evaluation phase[267]:

6.5.1 Evaluating data mining results

And in reference to chapter 2.4.5, the study investigates the generated models and determine the knowledge disovered from them.

In addition, the test results is used to determine the prediction accuracy of the generated models.

From an evaluation point of view, the clustering models which have been generated after the application of SimpleKMeans algorithms: 5 clustering on 2010-2013 Grade 12 examinations results are compared with with each other and possible causes of strength and weakness identified.

[267] (H. Du 2010)

Furthermore, the models generated during clustering focuses at, amongst the generated items, the following areas:

How the instances are spread out during the generation of the models.

What similarities and differences are there between clusters within the same data sets.

Interpretation of clusters in relation to average provincial performance rates.

6.6 Conclusion

With reference to chapter 2.4, this study is guided by the use of the CRISP-DM methodology, an industrial-standard mining methodology which ensures that all the stages in the mining process are followed and adhered to. Further to this, the use of the freely available software for data mining called WEKA makes it easier to produce results that are not only credible but internationally recognised as the software is widely used. The next chapter gives a detailed account of the output results obtained from the application of data mining processes on the 2010-2013 Grade 12 Motheo District Grade 12 results from FSDoE's examinations and assessment directorate.

Chapter 7 *Data modelling and evaluation*

This chapter presents the output of both the data modelling and evaluation phases of the CRISP-DM methodology followed by this study. These results include the output of the C4.5 classifier as well as the *simpleKMeans* clustering algorithm that were used for the supervised and unsupervised learning tasks respectively. The classification models were trained using the training data sets as discussed in Chapter 6.4. These models were then tested against the test datasets for which the confusion matrices were used to determine the accuracy of each model. This was an iterative process and the final models that were analysed and interpreted are presented in this chapter. During the unsupervised learning task one data set containing all the instances was used per subject. The five cluster definitions that were generated for each dataset i.e. per subject are discussed in Section 7.2.

7.1 Knowledge discovery by the C4.5 classifier

During the Classification task, 4 models were generated using 4 different sets of data as described in section 6.4. These data sets were:

English Home Language (GR12 2010-2013)

English First Additional Language (GR12 2010-2013)

Mathematics (GR12 2010-2013)

Mathematical Literacy (GR12 2010-2013)

The knowledge discovered during each of these classification tasks is discussed next.

7.1.1 Learner performance in English as a predictor of school performance

Two models were generated from English Language exam results. One from English Home Language (EHL) and one from csv data as illustrated in Figure 7.1.1 and 7.1.2.

Figure 7.1. 1 Output model generated by C4.5 classifier using English Home Language GR12 2010-2013).

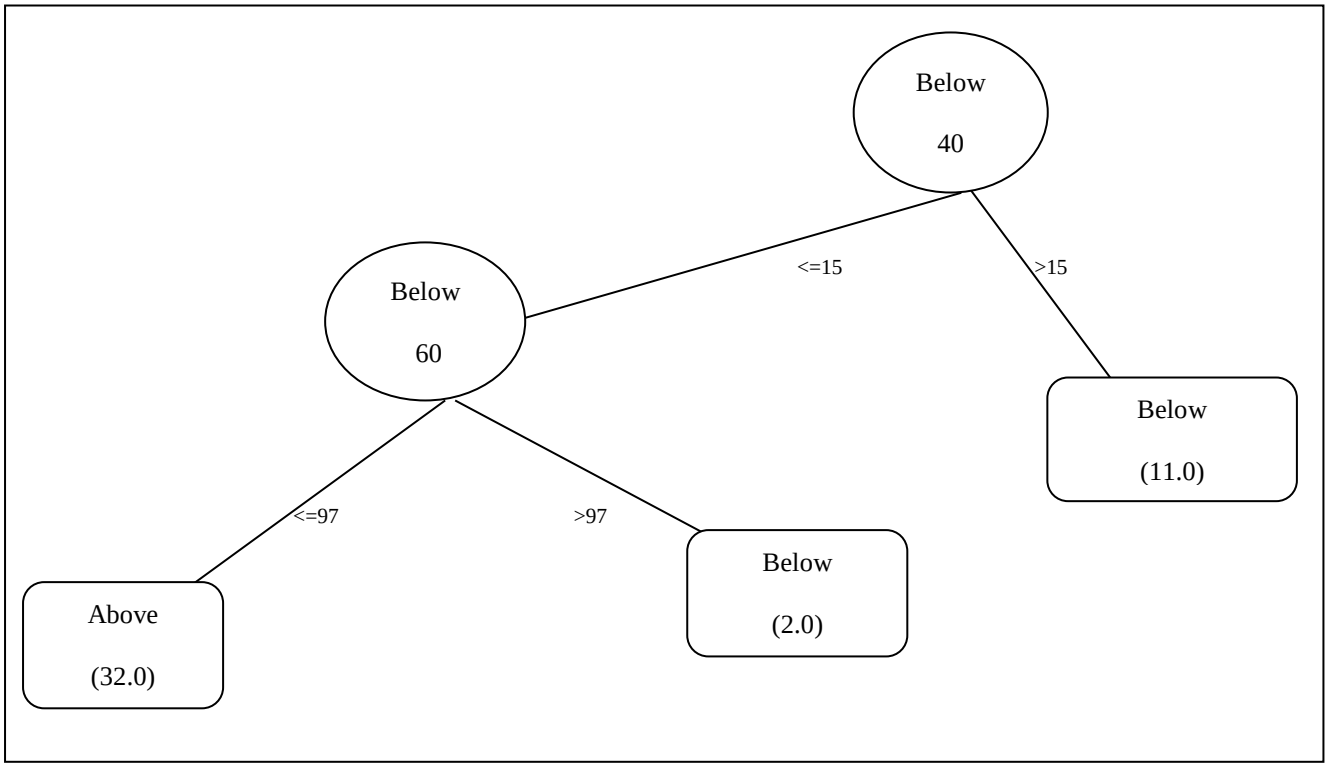
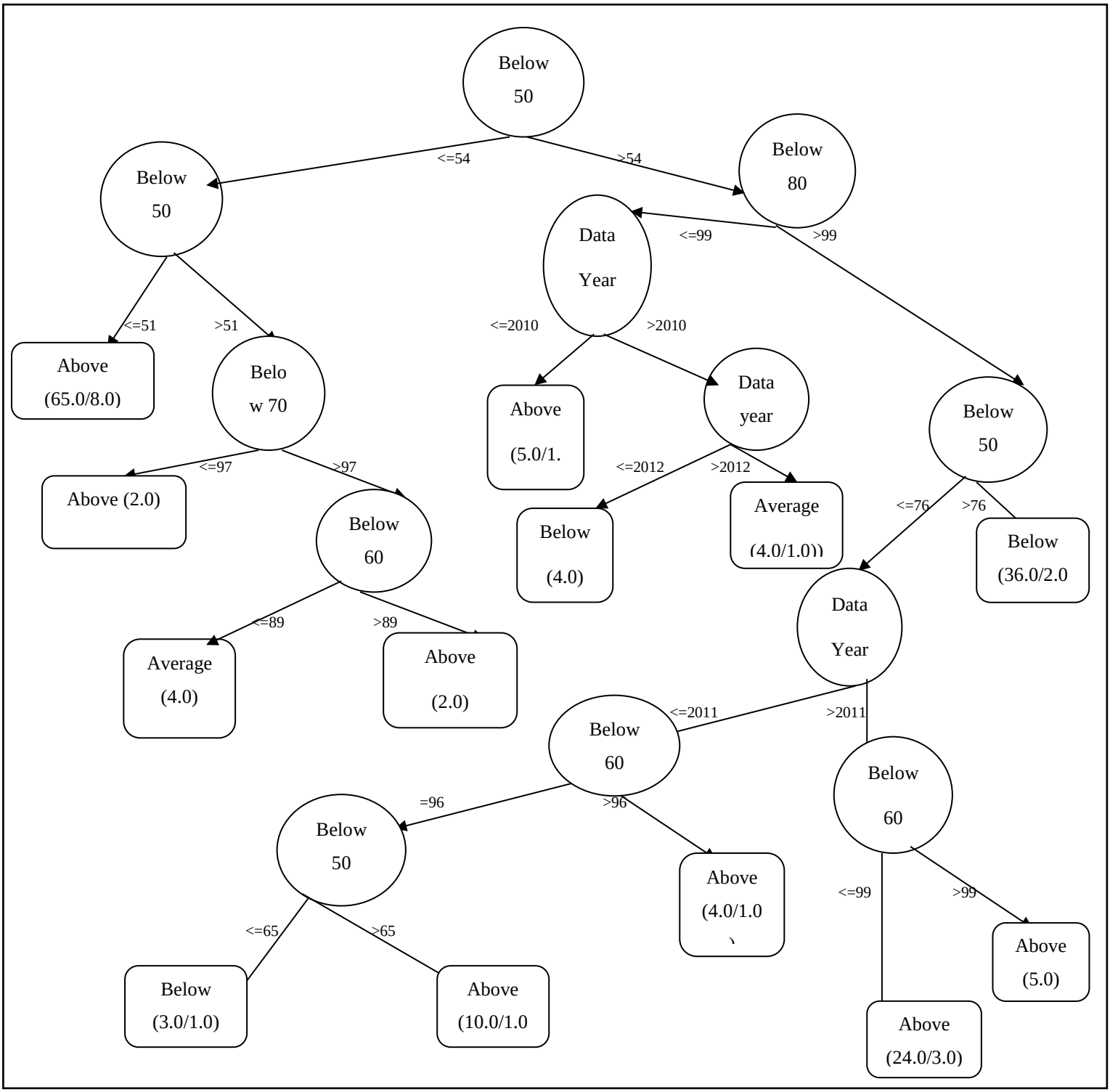


Figure 7.1. 2 Output model generated by C4.5 classifier using English First Additional Language (GR 12 2010-2013) dataset.



The number of instances in the two training datasets, 68 and 248, indicate the difference in the number of schools offering English Home Language versus English First Additional Language respectively in the Motheo District. During the 2010-2013 period, 11 (17%) of the schools

offered only English Home Language, 53 (83%) of the schools offered only English First Additional Language and 6 (10%) of the schools offered both English Home and English Additional Language. The knowledge discovered from the English Home Language model included the following rules:

In a given school in the Motheo District,

If more than 15% of the learners achieve a result below 40% for EHL

(performance lies within achievement level 1 and 2)

Then the school's overall performance will be in the Below category

Else

If at least 3% of the learners achieve a result above 60% for EHL

(performance lies within achievement levels 5 to 7)

Then school's overall performance will be in the Above category.

Due to the lack of Average class instances in the training dataset the model could only differentiate between Above and Below overall school performance. Secondly, because only 27% of the schools in the Motheo District offer EHL as a subject the model is applicable to only a small subset of schools. The model is however consistent over the 4 year period as no rule is based on the year of examination.

The accuracy of the model was determined using the test dataset which included 23 instances. The overall accuracy of the model was measured at 91%. As illustrated in the confusion matrix below.

a	b	← classified as
17	1	a = Above
1	4	b = Below

Although the model is a very accurate predictor of overall school performance based on EHL performance it is only applicable to a small number of schools within the district which limits it use.

However, in the Motheo District, 92% of the schools offered EFAL as a subject of choice during the period 2010 - 2013. Hence the training dataset included 268 instances and the generated model was more representative of the schools within the district.

The top node of the generated model was based on the Below 50 attribute value. If more than 54% of the learners perform below 50% in EFAL, the instances fall into the left branch of the decision tree. This branch of the decision tree does not include any nodes based on a 'Year' attribute value. However, in comparison the right-hand side of the decision (where more than 54% of the learners performed Below 50% in EFAL) include mostly nodes based on 'Year' values. When more that 54% of the learners within a school perform below 50% in EFAL, there are different set of rules, for the different years, to determine the schools overall performance. The 2010 period which is the first year selected for this study has an influencing role, and so are the 2011 and 2012 years. There are different set of rules associated with different years.

This could be interpreted as nothing out of the ordinary, or could be related to the point raised in parliament where Umalusi, had to defend the way they carried out standardisation as discussed in Chapter 1.1. The answers given by Umalusi were not only paradoxical, but also unconvincing. The fact that the generated rules are based on the value of the 'Year' attribute brings to the fore the primary question of this study, regarding the credibility of the Grade 12 examinations' data. It raises questions on whether the data was tampered with, and thus violating the General and Further Education and Training Quality Assurance Act No. 58 of 2001 and the Bill of Rights which are enshrined in the Constitution of the Republic of South Africa, Act no. 108 of 1996. Furthermore, were there adjustments made and if so how significant were they, as well as what methods were used to arrive at those percentages.

As indicated in section 1.2.2, the set objective of this study was to investigate the credibility of the Grade 12 results. This could be achieved by influencing decision makers who were either in line or against the norm set. The fact that underperforming schools' (Below classification value) is so strongly dependent on the year in which the EFAL examination was administered may challenge the decision makers to further analyse processes followed in the lead up to the publication of the final results during the four different periods. Since the model is primarily

based on the year the examination was administered it does not render itself suitable as a performance predictor for overall school performance based on a learners EFAL performance. The produced rules could only be used as a platform for introspection.

The accuracy of the model was determined using the test dataset which included 83 instances. The overall accuracy of the model was measured at 88%. As illustrated in the confusion matrix below.

a	b	c	←	classified as
31	3	0	|	a = Below
5	42	0	|	b = Above
2	0	0	|	c = Average

Like any DM model, the results may ever confirm or call for further inquiries on issues which show deviation from the norm and this model will assist in establishing the anomalies which could have taken place in the manipulation of Grade 12 examinations data during the specified period understudy.

7.1.2 Learner performance in Mathematics as a predictor of school performance

Two models were generated from Mathematics exam results. One from Mathematics and one from Mathematics Literacy data the classification rules extracted from these models are illustrated in Figure 7.2.1.1 and 7.2.1.2 for ease of interpretation purposes.

```
Below <= 86: Above (70.0/7.0)

 Below 60 > 86

|        Data Year <= 2010

|        |       Below 60 <= 99: Above (26.0/1.0)

|        |       Below 60 > 99: Below (10.0/1.0)

|        Data Year > 2010

|        |      Below 70 <= 98

|        |      |        Below 70 <= 95

|        |      |        |        Below 40 <= 62: Above (4.0/1.0)

|        |      |        |        Below 40 > 62: Below (3.0/1.0)

|        |      |        Below 70 > 95

|        |      |        |        Below 80 <= 99

|        |      |        |        |        Below 40 <= 64: Average (2.0)

|        |      |        |        |        Below 40 > 64: Below (4.0)

|        |      |        |        Below 80 <= 99: Above (5.0/2.0)

|        |      Below 70 > 98: Below (57.0/16.0)
```

Figure 7.1.2. 1 Classification rules generated by C4.5 classifier using Mathematics (GR 12 2010-2013) dataset.

```
Below 40 <= 34
|   Location = Township
|   |   Below 50 <= 65
|   |   |   Data Year 2012
|   |   |   |   Below 70 <= 90: Above (8.0)
|   |   |   |   Below 70 > 90
|   |   |   |   |   Below 70 <= 93: Below (7.0/2.0)
|   |   |   |   |   Below 70 > 93
|   |   |   |   |   |   Data Year <= 2011
|   |   |   |   |   |   |   Below 70 <= 97
|   |   |   |   |   |   |   |   Below 60 <= 76: Below (2.0)
|   |   |   |   |   |   |   |   Below 60 > 76: Above (17.0/ 5.0)
|   |   |   |   |   |   |   Below 70 > 97
|   |   |   |   |   |   |   |   Below 40 <= 23: Average (3.0/1.0)
|   |   |   |   |   |   |   |   Below 40 > 23: Above (2.0/1.0)
|   |   |   |   |   |   Data Year > 2011
|   |   |   |   |   |   |   Below 40 <=23 Above (3.0)
|   |   |   |   |   |   |   Below 40 > 23: Average (3.0/1.0)
|   |   |   Data Year > 2012: Above (10.0)
|   |   Below 50 > 65
|   |   |   Below 70 <= 96: Average (12.0/1.0)
|   |   |   Below 70 > 96
|   |   |   |   Data Year <= 2010: Average (2.0/1.0)
|   |   |   |   Data Year > 2010: Below (7.0/2.0)
|   Location = City
|   |   Below 40 <= 13: Above (44.0)
|   |   Below 40 > 13
|   |   |   Data Year <= 2012: Below (5.0/1.0)
|   |   |   Data Year > 2012 Above (2.0)
|   Location = Town
|   |   Below 80 <= 99: Above (7.0)
|   |   Below 80 > 99: Below (59.0/1.0)
Below 40 > 34: Below (59.0/11.0)
```

Figure 7.1.2. 2 Classification rules generated by C4.5 classifier using Mathematical Literacy (GR 12 2010-2013) dataset.

The number of instances for Mathematics and Mathematical Literacy training datasets are 181 and 191 respectively.

During the 2010-2013 periods, 68 (94%) Motheo District schools offered Mathematics and 72 (100%) schools offered Mathematical Literacy. Only 4 (6%) offered Mathematical Literacy whilst 68 (94%) offered both Mathematics and Mathematical Literacy.

The Mathematics model is similar to that of the EFAL model. The top node is based on the 'Below 60' attribute value. If less than 86% of the learners perform below 60% in Mathematics the school's overall performance is classified as an 'Above' average. However, if more than 86% of the learners perform below 60% in Mathematics the generated rules are once again based on the year attribute. What is interesting in this model is that only for 2010 a different set of rules were generated not for any of the other periods. This therefore indicates that the model has not given consistent rules which could be applied across the years except for 2010. This raises the same question that the members of the parliament posed to Umalusi regarding their standardisation processes. This could be due to some interference into Grade 12 examinations in 2010 which affected the credibility of the whole data set. Another possible influence could be policy changes during this time which influenced the examination results.

The accuracy of the model was determined using the test dataset which included 91 instances. The overall accuracy of the model was measured at 86%. As illustrated in the confusion matrix below.

a	b	c	←	classified as
32	3	1	|	a = Below
5	44	0	|	b = Above
2	2	2	|	c = Average

The confusion matrix has been able to predict with 86% accuracy, a schools overall performance based on the learners performance in Mathematics. The presence of 2010 'Year' attribute value, as a determining factor, may be considered as a starting point when conducting an investigation. The 86% accuracy may further be a call to decision makers to look at the tools that they use to ensure the credibility of the Mathematics Grade 12 data sets.

The model generated from the Mathematical Literacy data is unique in its own way as it is the only model that was generated where the location of the school is a major determining factor in the schools overall performance. Whether a school is located in a Township, City or Town has a marked influence on the schools at different performance levels. What can be derived

from this model is that the overall school performance of City, Township and Town schools can be predicted based on the learners' performance in Mathematics Literacy. This will enable decision makers in the FSDoE, when making use of this model with Grade 11 Mathematics Literacy data to predict school performance for Grade 12 and intervene in those schools with a predicted 'Below' classification.

The accuracy of the model was determined using the test dataset which included 97 instances. The overall accuracy of the model was measured at 86%. As illustrated in the confusion matrix below.

a	b	c	←	classified as
45	7	0	|	a = Above
3	38	0	|	b = Below
1	3	0	|	c = Average

The confusion matrix was able to predict the schools overall performance with a 86% accuracy based on the learners performance in Maths Literacy.

7.2 Knowledge Discovery by the SimpleKMeans Clustering algorithm

During the clustering task, 4 models were generated using 4 different sets of data as described in section 6.4. These data sets are:

English Home Language (GR12 2010-2013)
English First Additional Language (GR12 2010-2013)
Mathematics (GR12 2010-2013)
Mathematical Literacy (GR12 2010-2013)

The knowledge discovered during each of these classification tasks is discussed next.

7.2.1 Knowledge Discovery by the SimpleKMeans Clustering algorithm: English Language: 5 cluster

Two 5-cluster models were generated from English Language Grade 12 exam results for the periods 2010 to 2013. One is for English Home Language and another one for English First Additional Language data as illustrated in Tables 7.2.1.1 and 7.2.1.2.

Table 7.2.1 Knowledge Discovery by the SimpleKMeans clustering algorithm on English Home Language data set: 5 clusters.

Attribute	Full (Data) (68)	0 (16.0)	1 (13.0)	2 (8.0)	3 (17.0)	4 (14.0)
Data Year	y2010	y2011	y2010	y2011	y2010	y2013
Location	City	City	City	City	City	City
Below 40	9.7353	3.875	4.0769	0	31	1.4286
Below 50	39.3382	44.75	28.8462	4	76.2353	18.2857
Below 60	69.2059	84.375	61.4625	24.25	97.2353	50.7143
Below 70	89.6912	97.125	87.6923	70.375	99.9412	81.6429
Below 80	98.2059	99.8125	97.7692	96.375	100	95.6429
School performance	Above	Above	Above	Above	Below	Above

Table 7.2.2 Knowledge Discovery by the SimpleKMeans clustering algorithm on English First Additional Language data set: 5 clusters.

Attribute	Full Data (248)	0 (65.0)	1 (28.0)	2 (55.0)	3 (73.0)	4 (23.0)
Data Year	y2010	y2011	y2013	y2013	y2010	y2012
Location	Township	Township	City	Township	Township	City
Below 40	20.0242	11.1538	1.8214	9.4727	41.1096	4.3704
Below 50	53.6734	49.0154	9.8571	61.3273	80.7808	21.4444
Below 60	81.5242	84.8462	38.3571	91.3455	96.4384	57.963
Below 70	99.004	99.9231	93.7857	100	100	96.7407
School performance	Above	Above	Above	Below	Below	Above

The knowledge discovered from the 5-cluster English Language models has produced the following insights:

As shown in the two tables above, there are 68 and 248 instances involved in the generation of the English Home Language (EHL) and English First Additional Language (EFAL) - 5 clusters models respectively. The English Home language (EHL) model is evaluated first.

Cluster 2 has only 8 instances and does not take part in the evaluation of clusters as its impact cannot be comparable to the other clusters. The Data Year attribute y2010 is represented by two clusters: 1 and 3 with the latter (14 instances) chosen for evaluation as it has more instances than the former (13 instances). Cluster 4 (with 13 instances), though representing Data Year 2013, displays similar characteristics as those in cluster 0 (with 16 instances), which represents Data Year 2011. The results in cluster 0 are to be generalised as representing those in cluster 4. For this study, clusters 0 and 3 have been selected for evaluation in order to gain insights from the clusters.

In Cluster 0, 4% of the learners performed at the Below 40 category (which is between levels 1 and 2). In addition, 45% of the learners fall in the Below 50 performance category (which is between levels 1 and 3). The difference between the two categories is about 41%. In the Below 60 performance category (which is between levels 1 and 4), 84% of the learners fall in this category. The difference between the Below 50 and Below 60 categories is 39%. In the Below 70 performance category (which is between levels 1 and 5), 97% of the learners form part of this category. The difference between Below 60 and Below 70 is 14%.

In the below 80 (which is between 1 and 7), 99% of the learners perform in this category. The difference between the Below 70 and Below 80 categories is about 2%. In line with literature the upper categories (Below 70 and Below 80) are usually occupied by few number of learners and in the case of EHL there is a marked narrowing from both sides and the bulging middle which is where the majority of the learners are found. The performance of the schools in this cluster is Above the average.

This is in line with literature review which indicates that only 5% of the students always feature in the highest performance categories whilst the rest are either in the middle or below performance categories. There is a need to move the numbers to higher levels and that requires the implementation of proper intervention programmes similar to those suggested in literature Cluster 3 represents the controversial Data Year 2010 which has been mentioned extensively in chapter 1. Unlike other clusters, there are 31% of the learners who fall in the Below 40 category (levels 1 and 2). In the Below 50 (which is between levels 1 and 3) performance category, 72% of the learners belong into this category. There is a difference of 41% between the two categories. A bigger number of learner (over 70%) occupy these categories which indicates an uneven spread of performance and an indicator of few learners occupying the higher categories. In the Below 60 category (which is between levels 1 and 4), 97% of the learners belong to this category. There is a 25% difference between Below 50 and Below 60

categories In the Below 70 category (which is between 1 and 5 levels), 99% of the learners belong to this category. Between Below 60 and Below 70 there is a 2% difference confirming what literature says about few learners in the highest levels of performance.

In the Below 80 (which is between 1 and 7), 100% of the learners fall into this category. There is a 1% difference between the Below 70 and Below 80 categories and that conforms to a 5% of learners that literature claims are likely to occupy such a category. In this cluster, all the schools perform at Below average. This might be an indicator of what literature says about the interference that could have happened in the 2010 Grade 12 examinations data. This is further confirmed by the majority of the learners performing at between Below 40 and Below 50 (72%). This is a worrying result which might shed light or confirm that there may have been a manipulation of the Grade 12 data which would have compromised its credibility.

All clusters in the EHL model represent schools from the city. This might confirm the highest prevalence of EHL in the city schools. Another notable observation is that, except for cluster 3, there is an incremental increase (from smallest to the highest) in the percentage of learners who perform at different levels. The csv model is evaluated next.

There are 248 instances involved in the generation of the English First Additional Language (EFAL) 5 clusters model. Clusters 1 and 4 (which are represented by 28 and 23 instances respectively), has been left out of this analysis. Their instances have been spread too thinly and they do not give a true picture in relation to other clusters with many instances. Cluster 2 has similar characteristics to cluster 3 and therefore cluster 3 is chosen for evaluation. The evaluation results represent both clusters. Clusters 0 and 3 (which have 63 and 78 instances respectively) are analysed in order to give some insights.

In cluster 0, which represents the year 2011 Data Year, schools represented are from the township. About 11% of the learners fall into the Below 40 category (which is between levels 1 and 2). In addition, in the Below 50 category, (which is between levels 1 and 3) 49% of the learners fall into this category. The difference between the Below 40 and Below 50 categories is 38%. This means 38% of the learners occupy this category.

In the Below 60 category (which is between 1 and 4), 85% of the learners are found in this category. There is a 36% difference between the Below 50 and Below 60 category which further indicates that 36% of the learners are found in this category.

In the Below 70 (which is between levels 1 and 5), 98% of the learners fall into this category. The difference between the Below 60 and the Below 70 categories is 13%. This also shows, and as it is confirmed by literature, a marked decrease in number of learners as the levels of performance go up.

In the Below 80 (which is between level 1 and 7), 99.92% of the learners fall into this category. The difference between the Below 70 and Below 80 categories is 1.2%

This cluster represents schools which are at an Above average performance in EFAL. The majority of the learners occupy the Below 60 and Below 50 categories, confirming what literature says about the narrowing of numbers of learners as one move towards the highest levels of performance. It still maintains the long held belief that only 5% of the learners can perform at the highest levels.

In cluster 3, 41% of the learners occupy the Below 40 category whilst in the Below 50 category, there are 81% of the learners represented. The difference between these two categories is 40%. This shows a marked increase in the number of learners at the lower to middle levels, which may also be traced back to literature as representing the inconsistencies which happened in 2010 regarding the handling of Grade 12 examinations data.

In the Below 60 category, 96% of the learners are found in this level. The difference between Below 50 and Below 60 categories is 15%. There is a marked decrease and that is an indicator that fewer leaners occupy the upper levels of performance.

In the Below 70 (which is between levels 1 and 6) category, there are 99.8% of the learners represented which gives a difference of 3.8% to the Below 70 category. In the Below 80 category (which is between levels 1 and 7) 100% of the learners are in this category and the difference between it and Below 70 category is 0.2%. This category has very few learners performing and a worrying factor for the decision-makers who would like to see more learners performing at the highest levels. In overall, cluster 3 shows deviation from the norm which may be used as a starting point to look at how the data was adjusted or standardised at it was claimed by the quality assurance body Umalusi in chapter 1. In addition, the nature of governance in schools which may be compare to literature on the different types of loops that schools belong to.

7.2.2 Knowledge Discovery by the SimpleKMeans Clustering algorithm on Mathematics: 5 clusters

Two 5-cluster models were generated from Mathematics exam results. One from English Home Language and another from Mathematical Literacy data as illustrated in Tables 7.2.1 and 7.2.2.

Table 7.2.2. 3 Knowledge Discovery by the SimpleKMeans clustering algorithm on Mathematics data set: 5 clusters.

Attribute	Full (Data) (272)	0 (49.0)	1 (44.0)	2 (80.0)	3 (33.0)	4 (66.0)
Data Year	y2010	y2011	y2010	y2013	y2011	y2012
Location	Township	Township	City	Township	City	Township
Below 40	49.9816	65.6735	20.4091	46.0375	30.9697	72.3333
Below 50	69.761	87	35.0455	70.175	54.6667	87.1515
Below 60	83.9632	96.3469	56.3409	87.1	73.5758	94.5758
Below 70	92.7757	98.8571	75.7955	95.9	88.1818	98.0909
Below 80	97.2316	99.898	89.5455	98.75	96.4242	98.9394
School performance	Above	Below	Above	Above	Above	Below

Table 7.2.2. 4 Knowledge Discovery by the SimpleKMeans clustering algorithm on Mathematical Literacy data set: 5 clusters.

Attribute	Full (Data) (288)	0 (43.0)	1 (72.0)	2 (49.0)	3 (49.0)	4 (75.0)
Data Year	y2010	y2011	y2010	y2013	y2011	y2012
Location	Township	Township	City	Township	Township	Township
Below 40	24.5524	14.7209	3.8611	26.8776	31.2857	44.1333
Below 50	49.6181	42.2791	13.2083	59.6122	62.7143	73.6933
Below 60	71.2813	73.4419	32.4583	82.449	87.102	89.68
Below 70	85.3021	91.0465	55.3889	93.6531	97	97.6267
Below 80	93.7569	98.3023	77.7917	98.3469	99.5714	99.68
School performance	Above	Above	Above	Above	Below	Below

As indicated in the two tables above, there are 272 and 288 instances involved in the generation of the Mathematics and Mathematical Literacy- 5 clusters models respectively. The Mathematics model is evaluated first.

There are 272 instances representing the 5 Mathematics clusters. For evaluation purposes, clusters 1 and 3 (with 33 and 44 instances respectively) are not be included as they have few instances whose influence is very minimal in comparison to clusters 0, 2 and 4 (which have 49, 80 and 66 instances respectively).

Cluster 0, whose Data Year is 2011, is made up of schools from the townships. In the Below 40 category (which is between 1 and 2 levels), 66% of the learners are represented. This is a bigger percentage considering that it is the lower rung of the performance ladder. The Below 50 category (which is between levels 1 and 3) has 87% of the learners which gives a difference of 21 between the two lower levels.

The Below 60 category (which is between levels 1 and 4) has 96% of the learners represented which is a difference of 9% between the two categories (Below 50 and Below 60).

The below 70 category (which is between 1 and 5 levels) is representing 98.6% of the learners. The difference between the Below 60 and Below 70 categories is 2.6%. The below 80 category (which is between levels 1 and 7), has a 99.9% learner representation which is a 1.3% difference between the two categories (Below 70 and Below 80).

The schools in this cluster, are in the Below average performance category. In addition, the cluster represents many learners who have failed and who come from the township schools. This is indicated by many learners are in the lower categories of performance. This is an indicator or a challenge in terms of changing the state in which the schools are in, and, as literature says, either single-loop or double loop learning organisations. In addition, this also challenges schools to apply Bloom's Mastery Learning model and move away from the bloated lower poor performance to the higher best performance levels.

In cluster 2, whose Data Year is 2013, represents schools from the townships. In the Below 40 (which is between levels 1 and 2) category there are 46% of the learners who are represented. This is the biggest number considering the fact that it is the lowest category of performance. In the Below 50 category (which is between levels 1 and 3), there are 71% of the learners represented and the difference between the categories (Below 40 and Below 50) is 25%.

In the Below 60 category (which is between levels 1 and 4), there are 87% of the learners represented. The difference between the two categories (Below 50 and Below 60) is 16%. In the Below 70 category (which is between levels 1 and 5) there are 95.9% of the learners

represented. The difference between the Below 60 and Below 70 categories is 8.9%. The Below 80 category (which is between levels 1 and 7), there is a 98.8% learners representation. The differences between the two categories (Below 70 and Below 80) is 2.9%.

In line with literature, there are many learners who seem to struggle with the subject and occupy the lower levels of performance. There are also very few who are performing at the highest levels.

In cluster 4, with Data Year 2012, represents schools from the townships. In the Below 40 (which is between levels 1 and 2) category there is a 72% learner representation. In the Below 50 category (which is between levels 1 and 3), there is an 87% learner representation. The difference between the two categories (Below 40 and Below 50) is 15%.

In the Below 60 category (which is between levels 1 and 4), there is a 95% learner representation. The difference between the two categories (Below 50 and Below 60) is 8%. In the Below 70 category (which is between levels 1 and 5) there is a 98.09% learner representation. The difference between the two categories (Below 60 and Below 70) is 3.09%. The Below 80 category (which is between levels 1 and 7), has a learner representation of 98.94%. The difference between the two categories (Below 70 and Below 80) is 0.85%.

Like the other clusters, the trend is that schools in the township are struggling to perform at the highest levels when it comes to this subject. A bigger number of learners is at the two lower ends of the ladder whilst only a handful of learners scrape through and occupy the highest levels. The Mathematical Literacy model is evaluated next.

There are 288 instances representing the 5 Mathematical Literacy clusters. For evaluation purposes, both clusters 0 and 3 (with 43 and 49 instances respectively) represent Data Year 2011 and the cluster with fewer instances (cluster 0) are not be included as their influence is very minimal in comparison to other clusters. Clusters 1, 2, 3 and 4 (which have 72, 49, 49 and 75 instances respectively), are part of the evaluation process.

In cluster 1, which represents the Data Year 2010, represents schools in the City. The Below 40 category (which is between levels 1 and 2) has a 4% learner representation. The Below 50 category (which is between levels 1 and 3), has a 13% learner representation. The difference between the two categories is 9%.

The Below 60 category (which is between levels 1 and 4) has a 32.5% learner representation. The difference between the Below 50 and Below 60 categories is 19.5%.

In the Below 70 category (which is between levels 1 and 5), there is a 55.4% learner representation. The difference between the two categories (Below 60 and Below 70) is 22.9%.

In the Below 80 (which is between levels 1 and 6), there is a 77.8% learner representation. The difference between Below 70 and Below 80 category is 22.4%. the schools in this category are in the Above average performance which has many questions raised due to it representing the 2010 Grader 12 examinations data. This cluster has all the inconsistencies associated with the 2010 Grade 12 data in the FSDoE and amongst the notables is that learners only managed to reach the 77.8% level of performance (within the Below 80 category). This is the highest percentage which might mean the paper was so difficult that there were many failures. Even though adjustments in literature have been confirmed, the failure rate could have been so high that even after an adjustment no learner could perform above 78%.

In cluster 2, the Data Year is represented by 2013 data and the schools come from the townships. The Below 40 category (which is between levels 1 and 2) has a 27% learner representation. The Below 50 category (which is between levels 1 and 3), has a learner 59.6% representation. The difference between the two categories (Below 40 and 50) is 32.6%.

The Below 60 category (which is between levels 1 and 4), has a learner representation of 82.4% whilst the Below 70 (which is between the levels 1 and 5) has a 93.7% learner representation. The difference between the two categories (Below 60 and Below 70) is 11.3%. The Below 80 category (which is between 1 and 7) has a 98.3% learner representation. The difference between the two categories (Below 70 and Below 80) is 4.6%.

The cluster indicates that learner performance in Mathematical literacy by schools in the townships. As indicated in literature the lower levels have a 25% representation which means that there is still a challenge of a high number of failures in Mathematical Literacy. In addition, there are fewer learners towards the highest categories. The narrowing of numbers of learners confirms the 5% representation which literature has describes as common in subjects like Mathematical Literacy.

In cluster 3, the Data Year is represented by year 2011 and the schools are from the townships. In the Below 40 category (which is between levels 1 and 2), there is a 31.3% learner representation. In the Below 50 category (between levels 1 and 3), there is a 62.7% learner representation. The difference between the two categories (Below 40 and Below 50) is 31.4%.In the Below 60 category (which is between levels 1 and 4), has an 87.1% learner representation. The difference between the two categories (Below 50 and Below 60) is 24.4%.

In the Below 70 category (which is between levels 1 and 6), there is a 97% learner representation. The difference between the two categories (Below 60 and Below 70) is 9.9%.

The Below 80 category (which is between levels 1 and 7) has a 99.6% learner representation. The difference between the two categories (Below 70 and Below 80) is 2.6%.

In this cluster, the 2011 Grade 12 data indicates that schools in the township have a bigger representation at the lower level. By having 31% of learners, is a worrying figure and that means drastic measures need to be implemented to turn the schools around. This involves assisting them to be the best performing learning organisations and following Bloom's Mastery Learning model which can move more learners to the highest levels of performance.

Cluster 4, represents the 2012 Grade 12 examinations data and the schools come from the townships. In the Below 40 (which is between levels 1 and 2), there is a 44.1% learner presentation. In the Below 50 category (between levels 1 and 3), there is a 73.7%. The difference between the two categories (Below 40 and Below 50) is 29.6%.

In the Below 60 category (which is between levels 1 and 4), there is an 89.7% learner representation. The difference between the two categories (Below 50 and Below 60) is 16%. In the Below 70 category (between levels 1 and 5), has a 97.6% learner representation. The difference between the two categories (Below 60 and Below 70) is 7.9%. The Below 80 category (between levels 1 and 7), there is a 99.7% learner representation. The difference between the two categories (Below 70 and Below 80) is 2.1%.

In this cluster and typical of the township schools, the 2012 Grade 12 shows more learners who have failed (44.1%) and occupy the Below 40 levels. This then narrows dramatically the space to be occupied by the fewer learners from the middle levels to the highest levels.

7.3 Conclusion

The generated models have given some insights into the status of the selected Grade 12 data in the FSDoE during the period understudy. The controversy that surrounded the 2010 standardisation, for example, has been constantly linked to the models which were generated either through classification or clustering. In line with literature and the problem statement it has been observed that by applying DM tools and techniques the FSDoE will be able to make informed decisions by looking at the nature of the data and look for remedial actions in preparation for the future examinations, not only at Grade 12 but at all levels.

Chapter 8 *Conclusion and Recommendations*

8.1 Conclusion

In this chapter a consolidated report on the whole study is presented. It presents the concluding remarks on how the whole process links theory and practice with reference to the data obtained from the FSDoE's examinations and assessment Grade 12 data. The results that came from applying the DM tools and techniques are then used as reference points to make recommendations that will assist in further studies and in assisting the decision makers in the education sector to tackle the challenge of data generation, manipulation and analysis.

This study aimed at placing into practice the theory on the application of DM tools and techniques using Grade 12 examinations and assessment database. The primary question for the whole study is:

How can the application of data mining tools and techniques assist in establishing credible Grade 12 results?

The use of both theory and practice assisted in answering this pertinent question. The information gathered from theory confirms that data generation in organisations like FSDoE is increasing at an alarming rate. Such data which is stored in different databases is often left unattended where it becomes useless with little or no value at all. These tons of data are not only threatening to overwhelm the department but likely to lose value if left to gather dust. Without proper DM tools and techniques it would be impossible for institutions like the FSDoE to extract valuable information from such data. In order for the department to make a profit out of the data that they generate, they need to get a better understanding of what benefits would accrue from an application of such DM tools and techniques. The FSDoE's Grade 12 data, the main focus point of this study, has been studied and put into practice by following the universally accepted standards.

It is from the above primary question and its subsidiary questions that the selected data was subjected to the universally accepted CRISP-DM principles. It is through the application of this methodology that new knowledge was uncovered. The new knowledge was drawn from the identified patterns that different models displayed from the Grade 12's selected data. What make the methodology popular is its step by step approach and the ability to repeat the generation of models until the accepted results are obtained. As a new field of study DM

promises to be an important tool in assisting institutions like FSDoE make sense of the data that they generate on a daily basis.

Furthermore, it is important for the procurement departments in both private and public institutions to ensure that the DM software that is purchased for data understanding matches the needs of that particular organization. WEKA, though, freely available, has shown that it can assist the entire education sector in examining and assessing learners' assessment data. Growth in the number of the software companies that deal with DM mean institutions are likely to get the best return on their investment as variety will give them more choices to choose from.

As this study falls within the teaching and learning, it is important to link it with the relevant theories which assist in making sure that there is not only an improvement in the understanding of the data but also the environment where it is generated. The FSDoE's examinations and assessment data has shown that there are patterns which show loopholes in the teaching strategies that are used by different teachers. Bloom's Mastery Learning model has proven that if all the steps are followed correctly the students will not only perform at the highest levels but produce best data to work with and thus eliminate the need to do adjustment of marks. As indicated in the literature review, the teachers would need to allow all their students to successfully perform and achieve, at all times, at over 80% levels. It is not only possible to reach such a feat but also a change in mind-set and patience from the teachers.

The environment in which the schools find themselves can play an important role in improving the understanding of data that schools produce. The Grade 12 examinations data understanding, which is the focus of this study, will only improve if the schools are able to assess and position themselves as either single-loop or double-loop learning organisations. The schools have the ability to move from single-loop learning organisations where situations are allowed to stay as they are for a long time with no reviews in place to double-loop learning where schools are forever reviewing themselves and looking for strategies to take them to a higher levels. Schools, as evolving entities, have the ability to generate, better Grade 12 examinations data which will add quality and better understanding of the learners' needs.

On a national level, the presence of census influences not only the day to day running of the state but also the operations of the schools. It is from the census data that this study was able to link the identified patterns from the generated models with the results of the 2011 census data. The Grade 12 data in the FSDoE's examinations has proven to have a link to amongst things, the location of the school and furthermore the provision of both human and physical

resources. The socio-economic factors that the census has been able to provide indicate the existence of a direct link with the type of the Grade 12 data which was obtained from FSDoE's examinations database.

8.2 Recommendations

With reference to both literature review and the previous chapter on data analysis, the followoing recommendations focus on how the CRISP-DM software, Bloom's Mastery Learning model, schools as learning organisations and the results from the 2011 census can assist FSDoE'examinations an assessment section improves on handling data at a Grade 12 level. The recommendations have also taken into consideration the problem statement and the objectives of the study to make sense of the patterns that were generated after the application of DM tools and techniques on the FSDoE' s Grade 12 examinations and assessment database.

The study's main objective was to establish and identify patterns in the Grade 12 Grade 12 results as observed over a period of four years. These objectives are tied to the research question which seeks to answer if there is credibility to the FSDoE's examinations and assessment as observed over the 2010 to 2013 period. Literature in this study has shown that when clearly stated, the objectives can be applied on any assessment and examinations related project that the department would like to embark on.

The credibility of the Grade 12 examinations data would only be acceptable if, from the beginning there are clearly spelt out reasons for putting such data in a test. The decision makers differ in terms of portfolios and it is wise that there should be an assigned people within the department who will be able to craft and present the reasons and need for the examinations and assesment database to be assessed on a regular basis. The stated aims and objectives need to be clearly indicated and be accepted by all including the quality assurance body like Umalusi. It will be advisable to constantly reveiw the these objectives and making sure that they deliver and produce the tangible results. They must further be filtered down to lowere levels as that will ensure that there is a seamless movement of students in terms of performance throughout their schooling years.

Understanding the importance of applying DM in Grade 12 examinationa and assessment Grade 12 results should at all times be undertaken within the confines of law. This must start from the constitution to the specific rules and regulations within the department of education. Furthernore this means all stakeholders will have to be included in from the beginning in order to avoid being questioned at the last hurdle as it has been shown in chapter 1 where the

parliamentarians questioned the procuders that Umalusi follows when conducting the standardisation process.

As confirmed in the literature study, the whole process of DM requires enough time, AND this poses a challenge to the FSDoE to set aside time and acquire enough human and non-human resources. It has been indicated earlier that 50-80% of the whole DM project is spent on data preparation and this means this part will not only need committed staff but people with knowledge and skills about data handling and analysis . Both human and discovery systems working in tandem would assist in elimination of inconsistencies and inaccuracies that might arise during data preparation. Guided by the stated objectives it would be advisable for the FSDoE's examinations and assessment directorate to be in possession of the credible data that will lead to the realization of the stated objectives.

A constant consultation with all affected stakeholders like the schools, disctrict offices and the directorate would help to gather the correct data which usually differ from school to school. The importance of having staff members who are knowledgeable in the subject and use of the discovery system would help in making data understanding a seamless exercise. Tied to this is the aquiring of the correct computer and software DM tools that would suit the activities of the FSDoE's examinations and assessment directorate. Clearly spelt out specifications during the tendering process will ensure that the procurement directorate buys the correct equipment. If needs be, the prospective tenderers should be asked to demonstrate their DM software programmes prior to approval for purchasing by Supply chain management.

EMIS will have to take a leading role and ensure that from the beginning the tools that are used to collect data are credible and easy to operate. By having simple and user-friendly instruments will ensure that there will be a buy-in from the schools where this data is primarily found. It is also advisable that the infrastructure should be equally applied across the province irrespective of the socio-economic background of the learners in the schools.. The presence of enough man-power with the necessary skills will also ensure that, at all levels of data collection, trust, safety and credibility are mainatined at all times. It is also important that there should be a seamless network which connects all the schools databases to that in the districts and finally the Province. The chosen tools for collection and storage should follow imnternational standards in order to avoid compromisisng the credibility of the data which might cost a fortune later in the years.

Once the infrastructure has been put in place, it would be wise to make sure that data quality is sustained and that will occur if it is constantly cleaned and the missing data identified and

replaced in realtime. This is important because, as the literature indicates, data increases at a high rate, and if it is not captured at the correct time, it may be missed or incorrectly captured. In addition, as data differ in terms of form, it will also be important to know its nature and how it should be manipulated to suit the specific need s, for example, the application of DM tools and techniques in order to find specific patterns in a selected data set.

As it has been observed in this study, finding the accurate models is a tedious exercise, a clear action plan would need to be in place which would take the unpredictability of the output results into consideration. As indicated earlier, model-building is an iterative and interactive process that would need dedicated personell to specifically deal with the DM activities.

To find valuable patterns does not happen overnight. It is important that the software that is chosen to achieve the stated goals fit the job and is able to produce credible and acceptable results. As the market is awash with software which purports to provide solutions to DM applications. As literature has clearly indicated it will be wise to do a proper research on the correct software which can do the modelling from the data that the department deals with. It means there should always be value for money when justifying reasons for choosing a specific software. Since WEKA's DM software allows for repeated experimentation, it will help to generate as many models as possible before reaching a conclusion. The flexibility of the WEKA software also allows for the algorithms to be selected based on the identified neeeds and the pre-set objectives. The algorithms can be manipulated several times until the interesting patterns are generated.

It is by having proper models which are based on sound methodologies, like CRISP-DM that an acceptable analysis may be carried out in order to arrive at the desired conclusion. The models carry a lot of information which might at some point need a professional person to dissect and interprete them in order to make sense of what is hidden in the data under study. Although they were limited only to four subjects in one district, the generated models will save as a point of departure on how to maximise data manipulation strategies by using correct DM tools ans techniques.

Evaluation of the whole knowledge discovery processes will assist the department to look back at all the stages involved and identify those areas that need to be strengthend. Once convinced that all steps were followed to according to book, it would become easier to make a case about the results. They would need to consult on a continuous basis so as to ensure that the findings are not only correct but also understandable. This means convincing the decision makers that

the findings will benefit the organization and give it a competitive edge. The evaluation should make business sense and provide an answer to questions and challenges that the FSDoE might be facing when it comes to examinations and assessment of the Grade 12 learners.

It is important to note that the whole knowledge discovery does not promise to produce good or positive results only. The good ones need to be sustained and strengthened by providing the necessary support. If the school's good performance is consistent, the department's officials need to support them and ensure they move to the higher levels.

On the other hand, if there are any inconcistencies, they need to be identified, listed and tackled in a proper manner and a solution found. If for example, there are consistencies in the data results, such should be investgated even if it means going to the original raw data or looking at the previous years' perfomance of learners or average performance of the school. Such steps assist in closing up all loop-holes and ensuring that the stored data is credibility.

Although this study was focused on the application of DM on Grade 12 results in the FSDoE's examinations and assessment database, all the stages that Bloom developed for his Mastery learning can be subjected to DM algorithms. Beside the assessment stage which was part of this study and also part of the mastery learning, a number of algorithms may be used during the initial instruction, to obtain feedback and in the application of the corrective instruction. This also means DM can also work well in the classroom to help improve the whole teaching and learning process. Each stage will provide data that could be used to correct, improve or sustain the level of teaching and learning in the schools.

It should be noted that Bloom's Mastery Learning will succeeed only if all the stages are taken as a chain which need not be broken. This means that data production at Grade 12 level, means nothing if the classroom environment is not looked at. The schools which are struggling to perform consistently would need assistance from the specialists in different fields of study who will then make sure that teachers are fully empowered.

By accepting the differences in terms of capacity, teachers will be able to teach students according to their abilities. The students who are fast learners should be allowed to move at the fast pace as they quickly grasp the lessons as they are delivered. On the other hand, the teachers should also accommodate the learners who need an additional time in order to master the concepts and be able to apply them, for example, when tests are conducted. There should be an allowance for learners who need extra classes so that they manage and get the highest marks like their fellow students. By debunking the myth that a few learners can achieve at the highest

levels needs a lot of commitment from all stakeholders who will change the schools around and make them centres of excellence.

Based on the results from the study, the different schools display different levels in terms of the nature of loop learning that they display. In schools where there is consistent lack of improvement, is an indicator that they are operating at the zero-level loop learning and need a lot of attention and more support to undergo a complete overhaul in how they do their teaching and learning. The majority seem to display either or both first of second loop learning with the latter being the most acceptable of the two. Though not the ultimate, the schools at the first loop level need support to move to the second level. Even though there are few schools operating at the third level loop learning they may be used as models to influence the rest of the schools.

The root causes of such disparities in terms of gorvenace need to be revisited by categorising them and later making sure that they perform at the same level. It is a fact that others are already ahead of the pack whilst others are in a state of direpair. The identification of such levels means there should be resources and time set aside to make sure that the dysfunctional schools are turned into schools of excellence. It does not make sense to have well crafted policies when they are not practically implemented or delayed and the excuse being that there is not enough money. It is there for important that stakeholders are brought on board and be shown the state of affairs which will come from the collected data. By making use of the raw data will assit in making sure that we take informed decisions which will affect the future.

As it has been observed in this study, a number of schools across all locations showed results that were mainly concentrated in the middle categories, with a significant number still lagging behind and languishing in the lower categories. A few of the learners were in the upper categories. This coupled with the results of the 2011 census is an indication of many underlying factors which need to be taken into consideration.

Two notable factors that were indicated in the 2011 census report are the shrinking population in the Free State province and the steady increase in the private schooling. The influence of migration means the province will struggle to keep the same number of learners over a twelve year schooling period. The data which represent the progress of learners over a maximum 12 year period will not at all times be balancenced due to students leaving for other provinces. When it comes to an increase in private schooling the FSDoE will continue to face declining numbers of learners as parents who can afford take their children elsewhere. This, therefore

means, the FSDoE will have to look at the nature of the quality of service that they offer through the use of various DM tools and techniques.

The pattens that were identified after the application of DM tools and techniques have shown that there are areas that have proven to be strengths which the department can be produ of but the majority of the schools still reflect difference in more than one aspects which were reflected in the models that were generated from the models. In terms of decion-making FSDoE will be able to craft new strategies which will not only lead to an improvement in the teaching and learning in the province but also to the generation of models which will prove that the data is a true representation of what the different learners achieved in their respective subjects.

Bibliography

Adriaans, P. 1996. *Data Mining.* Harlow: Addison-Wesley.

Ahlemeyer-Stubbe, A. & Coleman, S. 2014 . *A Practical Guide to Data Mining for Business and Industry.* Somerset: John Wiley & Sons.

Argyris, C. 1991. *Harvard Business Review: teaching smart people how to learn.* [Online]
Available at: https://hbr.org/1991/05/teaching-smart-people-how-to-learn
[Accessed 27 April 2015].

Argyris, C., 2003. A life full of learning.. *Organization Studies,* 24(7), pp. 1178-1192.

Arlin, M. 1984. Time, Equality, and Mastery Learning.. *Review of Educational Research,* Vol. 54, No. 1 (Spring, 1984), pp. 65-86.), pp. 65-86.

Becerra-Fernandez , I. & Sabherwal, R. 2010. *Knowledge Management: systems and processes.*. New York: ME Sharpe.

Beikzadeh, M. R. 2008. Data Mining Applications in Higher Learning Institutions. *Informatics in Education,* 7(1), pp. 31-54.

Bensimon, E. M. 2005. Closing the Achievement Gap in Higher Education: An Organizational Learning Perspective. *New directions for higher education,* Volume 131, pp. 99-111.

Bernhardt, V. 2013. *Data analysis for continuous school improvement.*. New York: Taylor and Francis.

Bloom, B. S. 1968. *RELCV Topical papers and reprints No 1: Learning for Mastery.* [Online]
Available at: http://files.eric.ed.gov/fulltext/ED053419.pdf
[Accessed 31 07 2015].

Bousbia, N. & Belamri, I. 2014. Which contribution does EDM provide to computer-based learning environments?. In: A. Pena-Ayala, ed. *Educational Data Mining: Applications and Trends.* New York: Springer, p. 8.

Brown, J. 2012. The impact of the "Failure is not an Option Policy" on student grades".. *Perspectives in Learning: A Journal of the College of Education & Health Professions ,* 13(1), pp. 22-28.

Calders, T. & Pechenizyky, M. 2011. Introduction to the special section on educational Data Mining. *ACM SIGKDD Explorations Newsletter,* 13(2), pp. 3-6.

Campbell, M. 1999. *Knowledge Discovery in Deep Blue: A vast database of human experience can be used to direct a search.*
Available at: http://archive.computerhistory.org/projects/chess/related_materials/text/5-3%20and%2054.Knowledge_discovery_in_deep_blue/Knowledge_discovery_in_deep_blue.campbell-murray.1997.ACM.062303048.pdf
[Accessed 30 June 2014].

Chamberlain, D. D. 2012. Early History of SQL. *IEEE Annals of the history of computing,* 34(4), pp. 78-82.

Chen, C. & Wei, L. 2009. Data spread-based entropy clustering method using adaptive learning. *Expert Systems with Applications,* 36(10), pp. 12357-12361.

Chowdhury, S. 2009. Data mining tools and technologies for competitive business advantage. *European Journal of Management. ,* 9(2).

Cios, K. J. Pedrycz, Swiniarski, R.W. & Kurgan, L. 2007. *Data Mining: A Knowledge Discovery Approach.* New York: 2007.

Codd, E. F. 1970. A relational model of data for large shared data banks.. *Communications of the ACM,* 13(6), pp. 377-387.

Corr, L. & Stagnitto, J. 2013. *Agile Data Warehouse Design: Collaborative Dimensional Modeling from Whiteboard to Star Schema..* Leeds: DecisionOne.

Darwen, H. 2012. The Relational Model: Beginning of an Era. *IEEE Annals of the history of computing,* 34(4), pp. 30-37.

Darwen, H. 2014. *an-introduction-to-relational-database-theory-ebook,* UK: bookboon.com.

Davis, D. & Sorrell, J. 1995. *Mastery learning in public schools. Educational Psychology Interactive..* [Online]
Available at: http://chiron.valdosta.edu/whuitt/files/mastlear.html
[Accessed 08 March 2015].

Dibl, L. & Carbonel, Z. 2012. CLAG: an unsupervised non-hierarchical clustering algorithm handling biological data. *Bioinformatics,* 13(194), pp. 1-14.

Dringus, L. P. & Ellis. T. 2005. Using data mining as a strategy for assessing asynchronous discussion forums. *Computers & Education 45 (2005) 141–160. [Accessed 25 04 2015]*.

DU, H. 2010. *Data Mining Techniques and Applications: An Introduction.* Hampshire: Cengage.

Erdem, T. & Swait, J. 2004. Brand Credibility, Brand Consideration, and Choice. *Journal of Consumer Research*, 31(1), pp. 191-198.

Fayyad, U., Piatetsky-Shapiro. G. &. Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework. *KDD* , Volume 96, pp. 82-88.

Fayyad, U. & Stolorz, P. 1997. Data mining and KDD: Promise and challenges. *Future Generation Computer Systems*, Volume 13, pp. 99-115.

Fayyad, U. & Uthurusamy, R. 1996. *Data mining and knowledge discovery in databases.* Portland : AAAI.

Ferreira, R. 2012. *Writing a research proposal. In: complete your thesis or dissertation successfully: practical guidelines, edited by Maree K.*. Claremont: Juta.

Garcia , E., Romero, C., Venture, S., de Castro, C. & Calders, T. 2010. Association rule mining in learning management systems. In: C. Romero , S. Ventura, M. Pechenizkiy & S. J. D. Baker, eds. *Handbook of Educational Data Mining.* Boca Raton: CRC.

Garcia, E., Romero, C., Ventura, S. & de Castro, C. 2011. *EdITLib - The leading digital library dedicated to Education & Information Technology.*
Available at: http://www.editlib.org/p/53704/
[Accessed 12 October 2013].

Georges, A., Romme. L. & van. Witteloostuinjn. A. 1999. Circular organizing and triple loop learning. *Journal of Organizational Change Management*, 12(5), pp. 439-454.

Gilchrist, M., Mooers, D. L., Skrubbeltrang, G. & Vachon, F. 2012. Knowledge discovery in databases for competitive advantage.. *Journal of Management and Strategy*, 3(2), pp. 2-15.

Giudici, P. 2005. *Applied Data Mining: Statistical Methods for Business and Industry.* Sussex: Wiley.

Guruler, H. I. A. 2014. Modelling Student Performance in Higher education using data mining.. In: *Educational Data Mining: Applications and Trends.* New York: Springer, pp. 105-124.

Guruler, H., Istanbullu, A. & Karahasan, M. 2010. A new student performance analysing system using knowledge discovery in higher education databases. *Computers & Education,* 55(1), pp. 247-254.

Gxwati, N. I. 2011. *SUNScholar Research Repository.* [Online]
Available at: http://scholar.sun.ac.za/handle/10019.1/6487.
[Accessed 2012 May 17].

Hamalainen, W. & Vinni, M. 2011. Classifiers for educational data mining. . In: C. Romero, S. Ventura, M. Penchenizkiy & S. J. D. Baker, eds. *Handbook of Educational Data Mining.* Boca Raton: CRC PRESS.

Han, J. & Kamber, M. 2001. *Data Mining: Concepts and Techniques*.. San Francisco: Morgan Kaufmann.

Hofstee, E. 2006 . *Constructing a Good Dissertation: A Practical Guide to Finishing a Masters, MBA or PHD on Schedule.* Sandton: EPE.

Hornik, K., Buchta , C. & Zeileis, A. 2009. Open-source machine learning: R meets Weka. I. *Comput Stat* , Volume 24, p. 225–232.

Hossein, I. S. & Zahra, M. 2008. Application of Data Mining Tools to Predicate Completion Time of a Project. *World Academy of Science, Engineering and Technology,* Volume 42, pp. 204-209.

IBM, n.d. *IBM's 100 icons of progress: Relational Databases.* [Online]
Available at: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/reldb/
[Accessed 26 July 2015].

Inmon, W. H. 2000. What is a data warehouse? pp. 1-18.

Iranmanesh, S. H. & Mokhtari, Z. 2008. Application of Data Mining Tools to Predicate Completion Time of a Project. World Academy of Science, Engineering and Technology, (42):204-209.

Ivancevic, V., Knezevic, M. & Pusic, B. 2014. Adaptive testing in programming courses based on educational data mining techniques.. In: A. Pena-Ayala, ed. *Educational Data Mining: Applications and Trends.* New York: Springer.

Kay, J., Koprinska, I. & Yacef, K. 2011. Educational data mining to support group work in software development projects. In: Romero, C et al 2011. Handbook of Educational Data Mining. CRC Press: Boca Raton.

Khandar, P. & Dani, S. 2010. Knowledge Discovery and Sampling Techniques with Data Mining for identifying trends in data sets. *International Journal on Computer Science and Engineering.* , pp. 7-11.

Kimball, R., Ross, M., Thornthwaite, W., Mundy, J. & Becker, B. 2008. *The Data Warehouse Lifecycle Toolkit.* 2nd ed. Indianapolis: John Wiley & Sons.

Kok, F. H., n.d. *EMIS – On-Line.* [Online]
Available at: http://unpan1.un.org/intradoc/groups/public/documents/cpsi/unpan041424.pdf
[Accessed 27 July 2012].

Krzysztof , J. C., Pedrycz, W., Swiniarski, R. W. & Kurgan, L. A. 2007. *Data Mining: A knowledge discovery approach.*. New York: Springer.

Larose, D. T. 2005. *Discovering knowledge in data: an introduction to data mining.* New Jersey: : John Wiley & Sons.

Lovland, R. a. G. B. 2001. *You Raise Me Up.* [Sound Recording].

Luan, J. 2002. Data Mining and its applications in Higher Education. *New directions for Institutional Research,* Volume 113, pp. 17-36.

Maimon, O. & Rokach, L. 2005. *Data mining and knowledge discovery handbook.* Ramat-Aviv: Springer.

Maimon, O. & Rokach, L. 2010. *Data Mining and Knowledge Discovery Handbook.*. 2nd ed. Tel Aviv: Springer.

Mantz , Y. 2000. Developing a quality culture in higher education. *Tertiary Education and Management,*, 6(1), pp. 19-36.

McJones, P. 2009. *Oral History of Donald Chamberlin.* [Online]
Available at:
http://archive.computerhistory.org/resources/text/Oral_History/Chamberlin_Don/102702111.05.01.acc.pdf
[Accessed 30 August 2015].

Mehta, N. & Dang, S. 2011. Temporal Sequential Pattern in Data Mining tasks. *International Journal on Computer Science and Engineering*, 3(7), pp. 2674-2678.

Meyer, P. 1988. Defining and Measuring Credibility of Newspapers: Developing an Index. *Journalism & Mass Communication Quarterly*, Volume 65, pp. 567-574.

Michaud, P. 1997. Clustering techniques.. *Future Generation Computer Systems* , Volume 13, pp. 135-147.

Moradi, H., Moradi, A. & Kashani, L. 2014. Students' performance prediction using multi-channel decision fusion. In: A. Pena-Ayala, ed. *Educational Data Mining: Applications and Trends.* New York: Springer.

Norman, M. 2004. *Database Design Manual: using MySQL for Windows.*. London: Springer.

Pal , S. K. & Mitra. , P. 2004. *Pattern Recognition Algorithm for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing.*. London: CRC.

Pal, N. & Jain, L. 2004. *Advanced Techniques in Knowledge Discovery and Data Mining.*. London: Springer.

Parliamentary Monitoring Group. National Senior Certificate Examinations 2010: briefing by the department and Umalusi. 2010. *Parliamentary Monitoring Group.*. [Online]
Available at: https://pmg.org.za/committee-meeting
[Accessed 15 May 2012].

Pedder, D. & MacBeath. J. 2008. Organisational learning approaches to school leadership and management: teachers' values and perceptions of practice, School Effectiveness and School Improvement. *An International Journal of Research, Policy and Practice*, 19(2), pp. 207-224.

Pena-Ayala, A. 2014. *Educational Data Mining: Applications and Trends.*. New York: Springer.

Pena-Ayala, A. & Cardenas, L. 2014. How educational data mining empowers state policies to reform . In: A. Pena-Ayala, ed. *Educational Data Mining: Applications and Trends.* New York: Springer.

PhridviRaj, M. S. B. & GuruRaob, C. V. 2014. Data mining – past, present and future – a typical survey on data streams. *Procedia Technology*, Volume 12, p. 255 – 263.

Piety, P. J. 2013. *Assessing the Educational Data Movement.* New York: Teachers College Press.

Pollack, R. D. 2008. Data Mining: commin definitions, applications and misunderstandings. In: *Data Mining Methods and Applications.* Lawrence, K D; Kudyba, S; Klimberg, R K ed. New York: Auerbach.

Provost, F. & Fawcett, T. 2013. *Data Science for Business: what you need to know about data mining and data-analytic thinking.* 1 ed. Beijing: O'Reilly.

Qin, J. 1999. Discovering semantic patterns in bibliographically coupled documents. *LIBRARY TRENDS,* 48(1), pp. 109-132.

Ramakrishnan, R. & Gehrke, J. 2003. *Database Management Systems*, New York: McGraw-Hill.

Ranjan, J. & Khalil, S. 2008. Conceptual framework of data mining process in management education in India: an institutional perspective. *Information technology Journal,* 7(1), pp. 16-23.

Refaat, M. 2007. *Data Preparation for Data Mining using SAS.*. Armsterdam: Morgan Kaufmann.

Rokach, L. & Maimon, O. 2005. *Data Mining And Knowledge Discovery Handbook.* New York: Springer.

Rykiel, E. J. 1996. Testing ecological models: the meaning of validation.. *Ecological Modelling,* Volume 90, pp. 229-244.

Sankar, K. &. P. M. 2004. *Pattern Recognition Algorithm for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing.* London: Chapman & Hall/CRC.

Scribner, J. P., Cockrell, K. S., Cockrell, D. H. & Valentine, J. W. 1999. Creating Professional Communities in Schools Through Organizational Learning: An Evaluation of a School Improvement Process. *Educational Administration Quarterly Vol. 35, No. 1 (February 1999) 130-160.*, 35(1), pp. 130-160.

Seng, J. & Chen, T. 2010. An analytic approach to select data mining for business decision.. *Expert systems with applications ,* Volume 37, pp. 8042-8057.

Silvers, F. 2008. *Building and Maintaining a Data Warehouse.* London: CRC.

Smart Vision-Europe: Predictive Analytics for Smarter Business. 2015. *What is CRISP-DM Methodology?.* [Online]
Available at: http://www.sv-europe.com/crisp-dm-methodology/
[Accessed 20 April 2015].

Smith, H. 2005. What a BPMS is. *BP Trends,* Issue BP Trends, pp. 1-8.

South Africa. Constitution of the Republic of South Africa No 108 of 1996. 1996. *South African government.* [Online]
Available at: http://www.gov.za/sites/www.gov.za/files/images/a108-96.pdf
[Accessed 15 May 2012].

South Africa. Department of Basic Education. 2005. *Regulations for the conduct,administration and management of the assessment of the senior certificate.* [Online]
Available at:
http://www.gov.za/sites/www.gov.za/files/DoE_Regulations%20for%20the%20conduct,%20administration%20and%20management%20of%20the%20assessment%20of%20the%20senior%20certificate_11062010.pdf
[Accessed 14 April 2012].

South Africa. Department of Basic Education. 2012. *National Protocol for Assessment Grades R – 12.* [Online]
Available at:
http://www.education.gov.za/LinkClick.aspx?fileticket=BUB00bmth1I%3D&tabid=419&mid=2313
[Accessed 16 April 2014].

South Africa. Department of Education. 2005. *Natonal Education Information Policy,* Pretoria: Deaprtment of education.

South Africa. General and Further Education and Training Quality Assurance Act No 58 of 2001. 2001. *Umalusi.* [Online]
Available at: http://www.umalusi.org.za/docs/legislation/2001/actno58_2001.pdf
[Accessed 25 May 2012].

South Africa: Free State Department of Education, n.d. *Education Information Policy: Draft Data Quality Standards.* [Online]
Available at: http://www.education.fs.gov.za/
[Accessed 18 May 2012].

Srimani, P. & Patil, M. M. 2012. *A Classification Model for Edu-Mining.* Dubai, International Conference on Intelligent Computational Systems, pp. 7-8.

Statistics South Africa. 2012. *Statistics South Africa.2012. Census 2011 Statistical release – P0301.4.*. [Online]
Available at: http://www.statssa.gov.za/publications/P03014/P030142011.pdf
[Accessed 13 June 2014].

Sung, T., Chang, N. & Lee, G. 1999. Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction. *Journal of Management Information Systems,* 16(1), pp. 63-85.

Taras, M. 2005. Assessment summative and formative-some theoretical reflections.. *British Journal of Educational Studies,,* 53(4), pp. 466-478.

Teorey, T. J., Lightstone, S. S., Nadeau, T. & Jagadish, H. V. 2011. *Database Modeling and design: Logical design.*. 5th ed. Boston: Morgan Kaufman.

Trench, Andrew. 2012. *City-press.News24.* [Online]
Available at: http://www.news24.com/Archives/City-Press/The-painful-truth-about-our-exams-20150429
[Accessed 22 March 2012].

Trucano, M. 2006. Rethinking Education Management Information Syatems: Lessons from and options for less developed countries. *InfoDev,* Volume 6, p. 7.

Umalusi. Directives for certification. National Senior Certificate (schools), 2008. *Umalusi.* [Online]
Available at: http://umalusi.org.za/ur/publications/20081204directivesforcertificationnsc.pdf
[Accessed 18 05 2012].

Umalusi. Quality Assurance of Assessment: policies, directives, guidelines and requirements, 2006. *Umalusi.* [Online]
Available at: http://www.umalusi.org.za/docs/directives/2006/qaa_directives.pdf
[Accessed 05 May 2012].

Umalusi. The standardisation of the final examinations. 2007. *Umalusi.* [Online]
Available at: http://umalusi.org.za
[Accessed 23 May 2012].

Watson, R. T. 2006. *Data Management: Databases and Organizations.* New Jersey: John Wiley & Sons.

Witten, I. A. & Frank, E. 2000. *Data Mining: Practical Machine Learning Tools And Techniques* .. San Diego: Academic.

Witten, I. H., Frank, E. & Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* 3rd ed. Amsterdam: Morgan Kaufmann.

Zimmermann, H. J. 2006. Knowledge Management, Knowledge Discovery, and Dynamic Intelligent Data Mining.. *Cybernetics and Systems: An International Journal*, 37(6), pp. 509-531.