Measuring Academic Performance of Students in Higher Education Using Data Mining Techniques

By

Mohammad Alsuwaiket

A Doctoral Thesis

Submitted in partial fulfillment

of the requirements for the award of

Doctor of Philosophy of Loughborough University

Loughborough University

2018

ABSTRACT

Educational Data Mining (EDM) is a developing discipline, concerned with expanding the classical Data Mining (DM) methods and developing new methods for discovering the data that originate from educational systems. It aims to use those methods to achieve a logical understanding of students, and the educational environment they should have for better learning.

These data are characterized by their large size and randomness and this can make it difficult for educators to extract knowledge from these data. Additionally, knowledge extracted from data by means of counting the occurrence of certain events is not always reliable, since the counting process sometimes does not take into consideration other factors and parameters that could affect the extracted knowledge.

Student attendance in Higher Education has always been dealt with in a classical way, i.e. educators rely on counting the occurrence of attendance or absence building their knowledge about students as well as modules based on this count. This method is neither credible nor does it necessarily provide a real indication of a student's performance.

On other hand, the choice of an effective student assessment method is an issue of interest in Higher Education. Various studies (Romero, et al., 2010) have shown that students tend to get higher marks when assessed through coursework-based assessment methods - which include either modules that are

fully assessed through coursework or a mixture of coursework and examinations – than assessed by examination alone. There are a large number of Educational Data Mining (EDM) studies that pre-processed data through the conventional Data Mining processes including the data preparation process, but they are using transcript data as it stands without looking at examination and coursework results weighting which could affect prediction accuracy.

This thesis explores the above problems and tries to formulate the extracted knowledge in a way that guarantees achieving accurate and credible results. Student attendance data, gathered from the educational system, were first cleaned in order to remove any randomness and noise, then various attributes were studied so as to highlight the most significant ones that affect the real attendance of students. The next step was to derive an equation that measures the Student Attendance's Credibility (SAC) considering the attributes chosen in the previous step. The reliability of the newly developed measure was then evaluated in order to examine its consistency.

In term of transcripts data, this thesis proposes a different data preparation process through investigating more than 230,000 student records in order to prepare students' marks based on the assessment methods of enrolled modules. The data have been processed through different stages in order to extract a categorical factor through which students' module marks are refined during the data preparation process. The results of this work show that students' final marks should not be isolated from the nature of the enrolled

module's assessment methods; rather they must be investigated thoroughly and considered during EDM's data pre-processing phases. More generally, it is concluded that Educational Data should not be prepared in the same way as exist data due to the differences such as sources of data, applications, and types of errors in them. Therefore, an attribute, Coursework Assessment Ratio (CAR), is proposed to use in order to take the different modules' assessment methods into account while preparing student transcript data.

The effect of CAR and SAC on prediction process using data mining classification techniques such as Random Forest, Artificial Neural Networks and k-Nears Neighbors have been investigated. The results were generated by applying the DM techniques on our data set and evaluated by measuring the statistical differences between Classification Accuracy (CA) and Root Mean Square Error (RMSE) of all models. Comprehensive evaluation has been carried out for all results in the experiments to compare all DM techniques results, and it has been found that Random forest (RF) has the highest CA and lowest RMSE. The importance of SAC and CAR in increasing the prediction accuracy has been proved in Chapter 5.

Finally, the results have been compared with previous studies that predicted students' final marks, based on students' marks at earlier stages of their study. The comparisons have taken into consideration similar data and attributes, whilst first excluding average CAR and SAC and secondly by including them, and then measuring the prediction accuracy between both. The aim of this comparison is to ensure that the new preparation process stage will positively

affect the final results.

ACKNOWLEDGEMENTS

First praise is to Allah, the Almighty on whom ultimately we depend for sustenance and guidance. Second, special thanks should go to my advisors Dr. Christian Dawson and Firat Batmaz whose insightful remarks; invaluable comments and revision of my thesis are much appreciated. Their constant support and guidance have always helped me refine my work.

Special thanks go to my beloved parent Abdullah Al Suwaiket and Nora Al shebel who with all their means and sacrifices have provided me with endless love and all necessities. Furthermore, I am grateful to my adorable brother Mohannad who with his affection and laughter always helped me to look at the bright side of life.

It is my privilege to extend my deep appreciation and sincere gratitude to my beloved wife Manar Al Molouhe for her immense support, endless patience and all the sacrifices that she made on my behalf during these years. Lastly and most importantly, I'd like to express my endless and unconditional love for my daughter Noor who became a source of inspiration, positive energy and strength for me through difficult times.

Finally, I would like to thank my Country, Kingdom of Saudi Arabia for their help and their financial support are more than appreciated.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Artificial Neural Networks…….	ANN
Area Under the Curve…….	AUC
Association Rules Mining in Novel Genetic Algorithm.	ARMNGA
Coursework Weighting……..	CWW
Coursework Assessment Ratio……	CAR
Computer Based Environment ……...	CBE
Classification Accuracy……..	CA
Class Test Grade…….	CTG
Cross Industry Standard Process for Data Mining……..	CRISP-DM
Cumulative Grade Point Average……	CGPA
Data Mining ……	DM
Decision Trees…….	DT
Educational Data Mining ……	EDM
Exam Weighting……..	EXW
Genetic Algorithm……...	GA
Grade Point Average……...	GPA
Higher Education …….	HE
k-Nearest Neighbors……	kNN
Knowledge Discovery in Databases……	KDD
Learning Management Systems……..	LMS

Logistic Regression…….	LR
Massive Open Online Courses…….……..	MOOC
Mean Square Error……..	MSE
Module Mark……...	MM
Random Forest…….	RF
Refined Module Mark…….	RMM
Root Mean Square Error…….	RMSE
Student Attendance Credibility……...	SAC
Support Vector Machine…….	SVM
Virtual Learning Environment…….……...	VLE

CHAPTER ONE: INTRODUCTION

1.1 Introduction:

Students' academic performance has been an issue of interest by researchers and institutes since the earliest times of education management. Student performance has been defined in terms of students' Grade Point Average (GPA), their individual test grades, their enrolled single module success, or other forms that can be measured at specific times. A detailed definition of students' academic performance and its forms, is described in Chapter 2. In addition, predicting student performance has been an important area of investigation since the emergence of Educational Data Mining (EDM) as it can lead to improving the outcomes of Higher Education (HE), since predicting students' future academic performance may help direct students towards their strength areas and prevent them from continuing their performance retraction in certain areas.

What is also considered a vital issue, is determining the factors that may play a major role in improving the performance of HE students. These factors can be psychological, physiological, personal, or other factors such as attendance, grades, or the number of hours they spend studying.

Not surprisingly, it is being argued that student attendance in schools and HE institutes is correlated with students' achievements and performance. It has been demonstrated statistically, and through other methods, that student

attendance affects students' overall performance (Ahmed et al., 2013; Brijesh and Saurabh, 2011).

In order to consider student attendance as a criterion for estimating students' performance, student attendance data should be represented in a form that reflects the real weight of student attendance, rather than the number of attendance occurrences. In fact, previous studies that considered students' attendance have either considered the number of attendance occurrences as it is, or used its average value.

Furthermore, other indicators that have an impact on students' academic performance have been taken into consideration, particularly student marks. Past student marks (in any form), such as single exam marks, single module marks, or Grade Point Average (GPA), have been considered as inputs for predicting students' future performance in numerous studies. Nonetheless, after reviewing those studies, it appears that student marks have not reflected the enrolled module's assessment method. In other words, marks of students attending modules with different assessment methods should not be treated in the same way.

The above argument has formed, based on the fact that during the last few decades, the use of coursework-based module assessments has increased in the UK and other countries due to various educational arguments. Additionally, it appears that students prefer the assessment to be based on either coursework alone or a mixture of both coursework and exams because these types of assessments tend to yield higher marks than exam-based assessment alone

(Richardson, 2015). The increased adoption of coursework-based assessment has contributed to an increase over time in the marks on individual modules and in the proportion of good degrees across entire programmes (Bridges et al., 2002).

Accurate student assessment is an issue of concern in HE, especially as it is affecting students' performance, with students' grades often affected by the assessment methods of the enrolled modules. Changes in the use of different assessment methods have given rise to an increasing number of universities that are shifting from traditional exam-based to coursework-based (Heywood, 2000). Coursework-based assessment methods differ from exam-based assessment methods where knowledge or skill is tested for a very specific period of time. It has been widely acknowledged that the chosen assessment method will determine the style and content of student learning and skill acquisition (Heywood, 2000).

The choice of choosing the student attendance data and the module assessment method (coursework-based assessment, exam-based assessment or both), is due to the impact of these factors on students' academic performance, as mentioned in different studies.

There are numerous factors that may affect students' academic performance, as shown in the literature, such as student attendance, their background, their enrolled module's assessment methods, and many other factors, which are described in Chapter 2 in more detail. However, this thesis aims to combine

the pre-processed and modified data of students' attendance and marks, as a framework for predicting their performance.

After preparing data for students' attendance and marks, different Data Mining (DM) techniques will be applied in order to predict students' academic performance in terms of a final year's average.

1.2 Problem Statement:

In HE institutes, knowledge extracted from data, by means of counting the occurrence of certain events, is not always reliable, since the counting process sometimes does not take into consideration other factors and attributes that could affect the extracted knowledge.

As an example, student attendance in HE has always been dealt with in a traditional way, i.e. educators rely on counting the occurrence of attendances or absences building their knowledge about students, as well as modules, based on this count. This method can be unreliable and can neither provide a real indication of students' attendance nor can it assist in estimating any further measures (i.e. students' performance). Previous studies were concerned with students' attendance for estimating their performance without investigating the various factors that may affect the credibility of students' attendance itself. In other words, it appears that no studies have considered looking at other methods of recording student attendance, rather than relying on its traditional form (i.e. number of attendance occurrences) (Brijesh et al., 2011).

Student transcripts in HE have always been dealt with in a traditional way, i.e. educators rely on building their knowledge about students' performance based on students' marks (whether numerical or ordinal), assuming that all marks are the result of assessing students with the same assessment method. Student assessment methods in HE can be generally divided into two main categories: i) exam-based assessment, which includes different forms such as closed and open book examinations, essay-type exams, multiple choice exams; and ii) coursework-based assessments, which include research projects, assignments, and reports.

While different studies have shown that students tend to gain higher marks from coursework-based assignments than they do from examinations, (Chansarkar and Raut-Roy, 1987) studies have also found that combining exam-based and coursework-based assessment, as one assessment method, produced better average marks than examinations alone. For example, Gibbs and Lucas (2005) reported an analysis of marks on more than 1,700 modules at Oxford Polytechnic, where modules with 100% coursework had an average mark 3.5% higher than modules with 100% examinations, and there were three times as many failed students on modules where there were only examinations.

1.3 Research Aim:

Predicting students' academic performance in HE considering more accurate and credible students' academic data, using different DM techniques based on students' attendance and transcript data.

1.4 Research Objectives:

1. To investigate the factors of students' academic profiles, identifying those which have the biggest impact on their academic performance.
2. To use those factors in a way that guarantees more accuracy and credibility.
3. To show the differences between using data collected from students' academic records as it is, and using formulas that represent these data in a more accurate and credible form. Therefore, student attendance will not be recorded using conventional methods, where its value represents number of attendance/absence occurrences only, but also considering other factors such as the number of times instructors record attendance. Similarly, student marks will be looked at in relation to the module's assessment method.
4. To investigate existing studies regarding students' academic performance, including the research objectives, data source and Data Mining (DM) techniques which have been applied.
5. To show the importance of applying the DM classification techniques on HE students' data to predict and classify students' academic performance, while taking into consideration the accuracy of input data.
6. To evaluate the effect of the two new measures on the performance of the different data mining techniques.

1.5 Research Contributions:

The main contributions of this research are:

1. Modifying the Educational Data Mining (EDM) pre-processing stage to improve the quality of the model outcome.
2. Increasing the accuracy and credibility of the data collected from students' academic profiles - such as their attendance and transcripts - by replacing the traditional ways of handling such data.
3. Formulating students' transcript data in a way that guarantees the assessment methods of students' enrolled module reflects the original data. In other words, a contribution of this thesis is to consider the differences between various assessment methods and refine student marks based on these differences.
4. Formulating student attendance in a way that takes into consideration other factors that have an impact on students' attendance (such as number of times the instructors record attendance), rather than using the attendance data as it is (e.g. number of absences, poor or good attendance, etc.).
5. Based on above, this research also contributes towards predicting students' academic performance, considering more accurate and credible students' academic data, using different DM techniques.

1.6 Thesis Organization:

The remainder of this thesis is organized as follows:

Chapter Two reviews key concepts of Data Mining (DM), Educational Data Mining (EDM), factors that affect students' academic performance in HE, forms of academic performance, and other studies that applied DM's prediction technique, in order to draw conclusions between different factors and their impact on students' academic performance.

Chapter Three describes the methodology used to achieve the objectives of this research and discusses the proposed solutions that tackle the problems stated above.

Chapter Four consists of the analysis and experimental work, including preparing students' attendance and transcript data, formulating credibility of student attendance data and transcript data, and applying DM techniques.

Chapter Five presents the results and the evaluation of the work.

Chapter Six concludes the research outcomes and presents future work and recommendations.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction:

In this chapter, systematic literature review will be the main methodology that will be applied in this part of the thesis. A systematic literature review would map out related work studies and it allows to identifying gaps of knowledge; and it highlights the areas where additional research is required. However, all related work will be assessed and evaluated by arguing other studies comparing with this work. All related work has been classified into categories such as: objectives, the applied techniques and the data that used for each study.

Although analyzing educational data on a small scale is a relatively old practice, the advances in educational technology and the ability to collect large-scale educational data logs have led to investigating new methods and techniques for analyzing large-scale data collected from different educational sources. The focus on this particular area has translated into a new field that emerged from conventional Data Mining (DM) known as Educational Data Mining (EDM), which will be defined later in this chapter.

While this chapter tries to shed light on various professional views on EDM and its methodology, it also reviews different research that employed EDM in analyzing educational data, particularly higher educational data, to find correlations between different factors and HE students' performance.

This chapter reviews the history of EDM, starting with a definition and applications, and explores the use of EDM in predicting student performance

in HE, while emphasizing the methods and techniques used by researchers to draw conclusions about the relationship between initial data and predicted student performance.

2.2 Data Mining:

Data Mining (DM) is the exploration and analysis of large quantities of data in order to discover meaningful patterns (Tan, et al., 2004). DM can also be defined as "a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data" (Srikant and Agrawal, 1996).

DM has been seen as the process of knowledge discovery, usually referred to as Knowledge Discovery in Databases (KDD), a name given by Piatetsky-Shapiro (1989) in which there are many distinctions, but it basically consists of three main steps: pre-processing, DM and post-processing tasks.

In education systems, DM applications focus on how decision-making processes can be improved (Delaware et al, 2008) and more recently, data mining methodologies were used to enhance and evaluate HE processes such as the Cross-Industry Standard Process for DM, which is a DM process model that describes commonly used approaches that DM experts use to tackle problems (Delaware and Beikzadeh, 2004).

Since EDM is derived from conventional DM, the following section will go through the meaning and key concept of EDM and discuss its lifecycle in detail.

2.3 Educational Data Mining (EDM):

Educational Data Mining (EDM) is the application of DM to educational data. EDM is a learning analytics and quantitative observation method to understand how students react to the educational system. Its aim is to analyze educational data to resolve educational research issues. Nowadays, there is rapid growth in the educational field which has led to an increase in educational data, so mining of educational data has become a vital issue to understand student behavior during learning processes, or to understand student problems (Upadhyay and Katiyar, 2014).

EDM is concerned with developing different ways for exploring the unique types of data that come from the environment of education. That is, it is the process of applying DM to educational systems and data generated from educational settings (Romero et al., 2010).

Furthermore, EDM observes a process starting with finding the relationships between the data of the educational environment using DM techniques and validating the discovered relationships, so that uncertainty can be avoided. Following this, predictions can be made for the future on the basis of validated relationships in the learning environment and finally, it supports the decision making process with the help of predictions (Upadhyay and Katiyar, 2014).

2.3.1 Key Concept:

It is very important to outline the main, key concept of DM. In DM, there are various definitions that are commonly accepted by the DM community

(Kohavi and Provost, 1998). However, the following outlines the key aspects of DM, which should be defined before building any DM model.

Data: this term defines a set of recorded values that describe the state or behavior of an entity in line with a set of attributes or variables. These values vary within what is known as the attribute domains. These domains can be either continuous or categorical. Continuous values are also known as real values, i.e. derived from real numbers. On the other hand, categorical values could be either nominal (like names, colors, etc.) which have no relationship between them, or ordinal values which should have relationships between them (e.g. short, average, and tall) (Zaki and Meira, 2013).

Feature: is the instant representation of an attribute.

Record: while feature represents an instant of an attribute at a time, a vector of features represents a record.

Information: information is described as a set of patterns that underline a set of data (Witten and Frank, 2012). For example, a set of a computer science students' data can represent useful information ready to be extracted.

Information extraction process: is the process by which data sets are explored to identify underlying patters.

2.3.2 EDM Lifecycle:

EDM has a lifecycle whereby ordered phases have to take place in order to extract knowledge from gathered data, particularly educational data.

Different studies have categorized the EDM lifecycle into the following phases (Mariscal and Marbán, 2010) (Romero and Ventura, 2013).

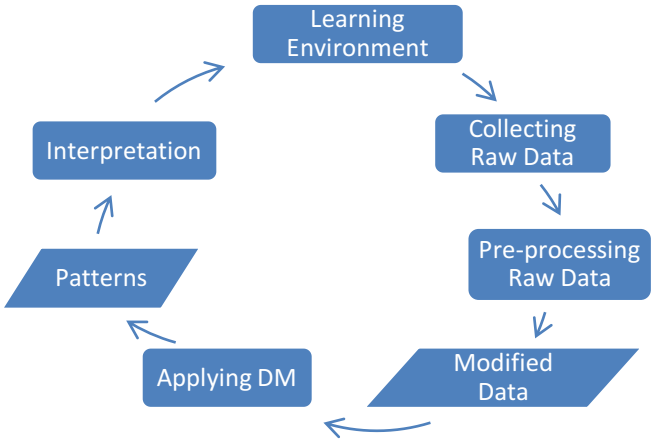


Figure 2.1 EDM Lifecycle (Mariscal and Marbán, 2010) (Romero and Ventura, 2013)

Figure 1 shows the main phases of the EDM lifecycle. The phases start with the learning educational environment system, the second phase is collecting raw data by gathering unmodified data that is relevant to the educational problem, and the third phase would be pre-processing the collected data and modifying it in order to then apply DM techniques. Finally, the results from the DM will be interpreted and evaluated. The following sections will look at these phases in depth.

2.3.2.1 Learning Environment:

In contrast to the conventional learning environment, where educators interact with students face-to-face, a Computer Based Environment (CBE) is referred to as "using computers in education to provide guidance, to instruct or to manage instructions to the student" (Romero, et. al., 2014). Currently, a CBE, equipped with web-based educational systems and artificial intelligence techniques, has encouraged the emergence of new educational systems such as: Intelligent Tutoring Systems (ITS) (Mostow, Beck, 2006), Learning Management Systems (LMS) (Brusilovsky and Peylo, 2003), Massive Open Online Courses (MOOC) (Suthers, et al., 2013), test and quiz systems (Brusilovsky, Miller, 1999), and other types of educational resources such as Wikis, blogs and forums.

In fact, the learning environment which is the focus of this thesis was applied to the traditional classroom-based environment, at institutions of Higher Education. EDM and its main methods will be applied, as they are used for predicting academic performance in a traditional setting, without using the Virtual Learning Environment (VLE).

2.3.2.2 Raw Data Collection:

Raw data collection is the process of gathering unmodified data that is relevant to the problem. That is, educators and developers tend to collect data related to a specific issue within the learning environment.

Educational data can be collected from various sources, such as: log files, quizzes and tests, and student portfolios (Zhu, et al., 2007). Data used in this research work consists of two main parts: Student Attendance Data and Student Transcript Data. While the former represents module-oriented data, which reflects the characteristics of student attendance at each module, the latter characterizes students' transcript data with regards to the enrolled modules' assessment methods, students' marks at each level of their study, their module marks, and other transcript-related attributes.

2.3.2.3 Data Pre-Processing:

Data pre-processing is where data is prepared in order for DM techniques to be applied. It is considered as the first and most important step in any DM process (Han and Kamber, 2006). Data pre-processing in EDM is differentiated from other areas as follows:

- Data collected from educational systems can be huge, particularly if it is gathered on a daily or even hourly basis.

- As mentioned previously, data includes the values that describe students; hence the number of attributes that describe students can be huge (personal, educational, psychological, ethnic, physiological, etc.)

The Pre-processing phase is the most important phase taken into consideration in this thesis, Chapter three will concentrate on the pre-processing phase. Within the data pre-processing phase, there are different tasks that can all or partially be applied to the data, before proceeding to

further phases of the EDM. This is summarized as follows:

- *Data Cleaning and Filtering:*

Data Cleaning is referred to as "detecting erroneous or irrelevant data and discarding it" (Han and Kamber, 2006). Missing data is a common type of error in educational data, and therefore there should be a mechanism by which missing data can be corrected with the most suitable values.

In contrast to other fields, such as motion detection, whereby missing data can be filled using mean values of both previous and posterior data cells (Salmeron-Majadas et al., 2013), in educational data this can lead to more erroneous data. Consequently, previous studies have used simple approaches for such problems, such as codifying missing/unspecified values by mapping incomplete values (Wettschereck, 2002).

On the other hand, Data Reduction involves reducing huge amounts of data, so it is smaller and more manageable. Since educational data can have hundreds or thousands of attributes per student (for example), educators or developers may not be interested in all of these attributes and hence, they can group a subset of the most relevant attributes and ignore the remaining ones.

For this reason, filtering can be used to choose only a specific subset of desired data (Mor, Minguillón, 2004). Removal of irrelevant data can be done by setting certain conditions on the data set; those attributes that fulfill the conditions remain in the subset, where the rest of the attributes can be

removed (Nilakant and Mitrovic, 2005).

- *Attribute Selection:*

Another task related to this thesis is the attribute selection task. Attribute selection is the process of selecting a subset of relevant attributes from the set of all available attributes (Liu, Motoda, 2007). For a more efficient data pre-processing phase, the most appropriate subset of attributes must be selected and those irrelevant or redundant ones should be ignored. Throughout the literature, several methods were used to solve the problem of attribute selection. In Márquez-Vera, et al. (2013) a ranking of several feature selection algorithms was used for identifying which features or attributes were most significant for predicting school failure.

In this thesis, an attribute selection task has been applied to select the relevant attributes from the available attributes, with irrelevant attributes being ignored.

- *Data Transformation:*

"Transformation is the data reduction and projection using dimensionality reduction techniques to condense the effective number of variables under consideration or to discover invariant representations of the data" (Fayyad et al., 1996a). This could simply mean to eliminate unwanted or highly correlated fields, to give the results validity.

Often, unmodified data can barely represent a certain behavior without transforming it into another form; for example, to divide a set of students into three groups based on their marks, marks should be transformed from a percentage (0%-100%) into an ordinal from {Poor (below 40%), Average (between 40% and 70%), Good (above 70%)}. In the above example, the process has produced a new attribute that was not in the original data.

Some examples of data transformation are: data normalization, discretization, derivation, and format conversion (Han, Kamber, 2006). Based on this, one of the main objectives of this thesis is to transform the student transcript and attendance data into new forms which will ensure this data can be linked to the attributes that effect this data.

2.3.2.4 Applying DM Techniques:

Also referred to in other contexts as modeling, it is the process of applying clustering, classification, predication, or association analysis techniques on modified educational data, in order to identify patterns and understandable rules.

Choosing the suitable DM technique relies on the problem itself; for example, in order to divide a set of students into groups based on their behavior in class, it is better to choose clustering algorithms. Additionally, algorithms within the same category differ, based on different factors such as size of data, or whether supervised or unsupervised learning is required.

Although traditional DM techniques were applied on educational data successfully (Romero and Ventura, 2013), educational systems have special characteristics that require a different treatment of the mining problem, in other words, some specific DM techniques are needed to address learning and other data about learners.

Currently, there are plenty of out-of-the-box tools that can apply several DM techniques on data, including, Weka, Orange, RapidMiner, ImageJ, and many more open-source or commercial tools that have capabilities from the pre-processing of data, DM techniques, to post-processing of knowledge.

2.3.2.5 Interpretation:

As important as the data pre-processing phase is, the interpretation of results, sometimes referred to as part of the post-processing phase of EDM (Bruha, Famili, 2000), is very important, since this will inform decisions about how to improve the educational environment or system (Ueno, 2004).

To interpret the data, knowledge extracted by applying DM techniques on educational data is simplified. This simplification process includes evaluating extracted knowledge, visualizing it (Cox, et al., 1997), or documenting it for end users (Bruha and Famili, 2000).

These steps have been commonly applied in both DM and EDM in recent times (Mariscal and Marbán, 2010), (Romero and Ventura, 2013), with no improvements made to this lifecycle, despite advances in computer manufacturing during the last decade. For example, the data preparation

phase has often been dealt with as a straight forward set of steps, neglecting the fact that this phase may have loopbacks to refine and improve the results of this phase. Romero and Ventura (2013), Baradwaj and Pal (2011) and Rüdiger and Hipp. (2000), stated that data preparation is closely tied with data modeling, yet it is not stated that there could be loopbacks within the data preparation phase itself. This thesis will show how results of the data preparation process can be fed back to the data set, creating various new attributes that can refine the data before feeding it into the modeling phase.

It can be concluded from the previous section that, in order to extract knowledge from gathered data, particularly educational data, the EDM lifecycle should be well defined. There are many applications of DM, but it is important to focus on the ones that deal with educational settings. In the next section, EDM applications will be presented.

2.4 EDM Applications:

Since its beginning, EDM was employed to discover new knowledge based on students' data, in order to help validate or evaluate educational systems, to potentially improve some aspects of the quality of education and to lay the groundwork for a more effective learning process (Cristóbal and Sebastian, 2010).

EDM has multiple applications such as data analysis and visualization, outlier analysis, Undesirable Student Behavior's Detection, grouping students, Constructing Learning Content and finally Predicting Academic

Performance. In this section, all EDM applications will be introduced but in this thesis one of these applications (predicting academic performance) will be explored in detail. Different studies have described the most common applications of EDM as follows:

2.4.1 Analysis and Visualization of Data:

The objective of analyzing and visualizing students' data is to provide educators with a view on students' learning process, through highlighting useful information and support decision-making. Two main techniques are being used in this application: statistics and visualization of information.

For statistics analysis, the educators or course administrators can use statistical software such as SPSS or Minitab to increase the efficiency and management of such a task (Wu, Leung, 2002). On the other hand, information visualization uses graphic techniques to help people understand and analyze any type of data (Mazza, 2009).

For outlier analysis, Grubbs (2011) stated that, "an outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs" and helps in identifying system faults and frauds before they escalate with potentially disastrous consequences.

"Outlier analysis is primarily studied as an independent knowledge discovery process merely because outliers might be indicators of interesting events that have never been known before" (Grubbs, 2011). Despite the advances seen, many issues of outlier analysis are left open or not yet completely resolved.

Outlier analysis is an important data-mining task. It deserves more attention from data mining community. There are good outliers that provide useful information that can lead to the discovery of new knowledge and bad outliers that include noisy data points.

2.4.2 Undesirable Student Behavior's Detection:

This applies to students who may have certain problems or abnormal behaviours, such as low motivation, distraction and cheating tendencies. Defining undesirable behaviours correctly and clearly is vital for reducing behaviours with appropriate strategies (Aydın; 2000). Recently, EDM techniques have used to detect such students behaviours.

Various EDM techniques such as Clustering, Decision Tree and Random Forest techniques have been applied to students' online exams (Burlak, G., et al., 2006), (Vee, M.N., et al., 2006) and classification (Kotsiantis, S., et al., 2003), (Lykourentzou, I., et al., 2009) and can be used to detect such behaviours. all previous studies have applied different data mining techniques to detect students' undesirable behaviours However, the main objective was to decrease these behaviours and increase their academic performance.

2.4.3 Grouping Students:

Different students have different personalities, features, characteristics, etc. Therefore, the instructor or course developer may group those with similar characteristics in order to promote an effective group learning system. The EDM technique, largely involved with grouping students with similar

characteristics, is clustering (Romesburg, H.C., 2004). The authors have grouped students with similar characteristics by applying k-means clustering technique in order to promote an effective group learning system, however, the authors defined the association between grouping method and group outcomes in terms of students' learning experiences, perceived problems, group leadership satisfaction and group task satisfaction. Also they decided which group of learners amongst the intelligently grouped, randomly grouped and instructor grouped methods using GPA collaborates more effectively and performs better in an online group task.

2.4.4 Constructing Learning Content:

The objective of this application of EDM is to help instructors in constructing and developing learning content automatically. The role of DM in this application is through the DM techniques, clustering and naïve algorithms, used to construct personalized learning material by building a personalized web tutor tree (Tang, C., et al., 2000). Another example is using Rough Set theory and the Clustering Concept Hierarchy to construct e-learning FAQ retrieval infrastructures (Chiu, D.Y., et al., 2008).

2.4.5 Predicting Academic Performance:

Cristóbal and Sebastian (2010) stated that "the objective of prediction is to estimate the unknown value of a variable that describes the student." That is, prediction involves analyzing and investigating the past and present values of variables that describe a student, in order to provide an estimate of a certain

future output for that same student, such as the student's marks and student's attendance.

The prediction process involves estimating interesting (and sometimes unexpected) connections between the output (performance in this case) and various inputs. The literature review on predicting student performance in HE by using DM techniques will be presented in the next section.

In addition, an overview of the DM techniques that have been used to predict students' performance, will be provided, including how the prediction algorithm can be used to identify the most important attributes in student's data.

2.5 Predicting Students' Performance in Higher Education (HE):

While the list of EDM applications grow with time, this thesis emphasizes the applications related to predicting student performance in HE. This involves discussing the definitions and forms of student performance in HE.

Student performance can be defined as a measure of students' academic achievements in a fixed period of time, considering different factors that may affect this measure.

In Cristóbal and Sebastian (2010), student performance is defined as values that describe the students, which represent a knowledge mark. Alternatively, according to Quadri and Kalyankar (2010) students' academic performance is defined by students' cumulative grade point average (CGPA) upon

graduating.

2.5.1 Forms of Student Academic Performance:

Student academic performance, as defined in section 2.5, is a set of values that describe students, such as Student GPA, individual test grades, single module's success and measuring errors made by students. The following forms of the previous values are reviewed in the sections 2.5.1.1, 2.5.1.2, 2.5.1.3, 2.5.1.4, 2.5.1.5:

2.5.1.1 Student GPA:

Garey (1981) and David (1993) considered students' GPA as an indicator of their academic performance, and based on that, the authors investigated students' attendance rates and their effect on academic performance. Minaei-bidgoli, et al. (2003) compared different DM methods to predict student performance, represented by their final grade, based on features extracted from Moodle's logged data. More specifically, Bayesian networks have been used to predict a future graduate's cumulative GPA based on applicant background at the time of admission (Hien and Haddawy, 2007) (Al-Radaideh, et al, 2011). The main objectives of the previous studies were increasing the students' academic performance by predicting their future performance considering the GPA as an major factor that affect academic performance.

2.5.1.2 Individual Test Marks:

Gabrilson (2008) first used DM prediction techniques (classification, Predictive model, and Bayesian classification) to highlight the factors that affect students' test grade. Gabrilson (2008) then adjusted these factors to improve the performance of students' test marks, and so set students' individual test grades as indicators to student performance.

Fausett and Elwasif (1994) used Artificial Neural Networks (ANN) to predict academic performance from test marks (using back-propagation and counter-propagation). Other DM methods have been used to predict students' marks. Nebot, et al. (2006) applied fuzzy association rules for mark prediction in e-learning environments while Shangping and Ping (2008) used an approach for classifying students in order to predict students' final mark, based on features extracted from logged data in an education web-based system. Bayesian networks were also employed for this task. Authors take advantage of the Genetic Algorithm (GA) designed specifically for discovering association rules. They proposed a novel DM algorithm, called Association Rules Mining in Novel Genetic Algorithm (ARMNGA), compared to the fuzzy association rules algorithm. The ARMNGA algorithm avoids generating impossible candidates, and therefore is more efficient in terms of the execution time (Shangping and Ping, 2008).

Ayers and Junker (2006) have used the Bayesian networks to directly examine the alignment between tutoring tasks and assessment items and to use the Transfer model to build more efficient functions for predicting end-

of-year exam performance. This was achieved by looking at student activity with the online tutor with a sample of over 900 eighth-grade students, who used an online intelligent tutoring system. Authors have built an accurate model for the tutoring and exam data, to improve an apparent lower bound on mean absolute prediction error.

All reviewed literature in this section is using students' test marks records as it is and do not use formulas that represent more accurate forms of students' marks in such modules. In this thesis student marks will not be used as they stand, instead, a new method of re-evaluated student marks, which are more accurate, will be proposed.

2.5.1.3 Single Module's Success:

Sometimes, the values representing students' academic performance could be, not only numerical (GPA, grades, etc.), but also categorical indicators, i.e. pass versus fail. Hämäläinen and Vinni (2006) have carried out a comparison of machine learning methods to predict success in a module (either passed or failed) in Intelligent Tutoring Systems. Other comparisons of different DM algorithms are made to classify students (predict final marks) based on Moodle data (Romero, et al., 2008). Alternatively, Delgado, et al. (2006) had utilized neural networks' radial basis functions to predict students' pass or fail ratios from Moodle logs of the accesses from a group of 240 students of data processing of the University of Cordoba, registered in the subject Methodology and Programming Technology. The proposed model of this study shows that it is possible to predict whether those students will pass a

course. In an educational environment, 80% success the course could be interesting. The scope of this work is to propose a tool that helps the teacher pay more attention to those students that are more likely to fail.

Martinez (2001) has employed a different DM technique for this task, using a multiple regression technique for identifying variables that could predict success in college modules.

2.5.1.4. Measuring Errors Made by Students:

In contrast to measuring right answers and total scores, some studies employed different DM techniques to measure and predict errors that student make. Using feed-forward and back-propagation, Want and Mitrovic (2002) predicted the number of errors a student will make by proposing an intelligent problem-selection agent, which identifies the appropriate problem for a student in two stages. The first stage is predicting the number of errors the student will make on a set of problems, and the second stage is to decide a suitable problem for the student.

2.5.1.5. Students' Marks:

Obviously, extracting student-related knowledge that can be useful in terms of clustering or grouping students, predicting their future performance, or any other application, cannot be done without taking into consideration students' marks. In the literature, studies have considered the impact of this indicator in various ways. Gedeon and Turner (1993) predicted students' final grade based on partial grades from during the teaching session, using back-

propagation trained feed-forward neural network. Similarly, Fausett and Elwasif (1994) used two neural networks for general mapping problems, back-propagation and counter-propagation, which are trained to predict students' grades in Calculus I from placement test responses. Moreover, Pritchard and Warnakulasooriya (2005) used the probability of correct responses on the first attempt of an exam to predict students' final marks. Additionally, Osmanbegović and Suljić (2012) applied three supervised DM algorithms on the pre-operative assessment data (that included students' GPA as a variable) to predict success in a course (either pass or fail). This included, the data collected from the surveys conducted during the summer semester at the University of Tuzla, (Faculty of Economics) in the academic year 2010-2011, among first year students and the data taken during enrolment.

Yadav et al. (2012) used Class Test Grade (CTG) as a variable, which is the result of calculating the average of class tests per semester. They split the CTG into 3 classes: Poor (less than 40%), Average (greater than or equal to 40% and less than 60%), and Good (greater than or equal to 60%). Besides other variables, authors used CTG to compare three different DM decision tree algorithms. A similar method was used in Baradwaj and Pal (2011) to help identify the dropouts and students who need special attention, allowing the teacher to offer appropriate guidance and advice.

Many other studies have been using student marks (in their different forms) as an input variable for EDM processes since the beginnings of DM on

education data, albeit with the changes to students' assessment methods (Heywood, J., 2000).

According to the literature, it appears to be that these changes were not reflected in the conventional way of using student marks as an input variable in the EDM process. In fact, student marks should not be isolated from the module's assessment method. This means that student marks should be prepared during the data preparation phase, in a way that allows refining the marks based on the assessment methods of the module.

Therefore, this thesis also investigates the changes on the recent assessment methods and tries to reflect these changes in the input variables, particularly where student marks are concerned.

After reviewing the most common forms of students' academic performance, it is vital to review factors that affect these forms of performance.

2.5.2 Factors Affecting Student Performance

In the literature, students' academic performance can be affected by various factors. This section will review the most common factors that have an impact on students' academic performance in its different forms.

2.5.2.1 Student Attendance

It has been proven by studies highlighted in the literature reviewed in this thesis, that students' attendance affects their overall academic performance.

The earliest studies that tackled this problem were Breiman, et al., (1984), Garey (1981) and David (1993), who adopted student attendance rates as criterion and used Decision Trees (DT) and Naïve Bayes (NB) data mining techniques to conclude that lower attendance rates negatively affect students' GPA.

The study by Romer (1993) is considered as one of the major publications that explored the relationship between student attendance and student performance in exams. From a preliminary survey on the extent of absenteeism at US colleges, Romer (1993) found that, on average, about one third of the students attending undergraduate Economics courses, at a representative sample of American universities, were missing class on a "typical" day. Romer put this result on consideration and used attendance records at six meetings of one of his large size courses to provide quantitative evidence on the effect of attendance on students' performance in exams.

Similarly, in term of student performance, Hijazi and Naqvi (2006) were selected a sample of 300 students (225 males, 75 females) from a group of colleges in Punjab university at Pakistan. Their hypothesis stated was "Students' attitude towards attendance in class, hours spent studying on a daily basis after college, students' family income, students' mother's age and their mother's education, are significantly related to student performance." Bharadwaj and Pal. (2011) obtained university student data on attendance, class tests, seminar and assignment marks from the students' previous database, to predict their performance at the end of the semester.

Some other studies have concerned about the impact of class attendance on students learning performance such as (Njål Foldnes, 2017), (Anna Lukkarinen et al., 2016), the authors have investigated the relationship between class attendance and students' academic achievements in HE and they found that student attendance is positively and significantly related to performance.

By reviewing the literature that considers students' attendance as a criterion for predicting academic performance, it appears that all the studies have either represented attendance by a numerical value (i.e. number of attendance occurrences), (Mazza and Milani, 2004), by ordinal values (i.e. poor, average, or high attendance), (Yadav et al., 2012), or by boolean values (i.e. yes/ no attendance), (Quadri and Kalyankar, 2010).

It can be concluded that all the literature uses data collected from students' academic records as it stands and do not use formulas that represent more accurate and credible forms of this data. In this thesis student attendance will not be used in the conventional way where its value represents attendance/number of absence occurrences only, but also considers other factors, such as the number of times instructors recorded attendance. Moreover, a new method of measuring student attendance more accurately will be proposed in this thesis.

2.5.2.2 Other Factors:

Through the literature, it is clear that students' attendance and marks are not the only impacts on academic performance, but various other factors may affect students' academic performance. It should be mentioned here that other factors will not be used in this thesis for some reasons such as the limitation of data sources, the proposed factors and indicators are the most common and popular more than others regarding to the literature review. Moreover, there are other factors which it is not possible to explore in this thesis. However, a student's GPA and attendance data were found to be significant factors and will be taken as the main, significant factors in this thesis.

Hien and Haddawy (2007) tried to shed light on the relationship between students' background at the time of admission and their future graduate's cumulative GPA. Also, students' end-of-year exam performance was predicted using the following data:

- Student activity with online tutors (Ayers and Junker, 2006)
- Learning management system such as Moodle usage data (Romero et al., 2008), (Delgado, et al., 2006)
- Students' learning portfolios (Chen et al., 2007)
- Features extracted from logged data in an education web-based system (Shangping and Ping, 2008)
- Log and test scores in web-based instruction, using a multivariable regression model (Yu et al., 1999)

2.5.3 Student Assessment Methods

In the previous sections, the forms of student's academic performance and the factors that affect students' performance in HE have been presented and discussed. Student assessment methods will now be discussed in HE, especially in UK universities, and how these methods might be useful for predicting students' performance.

Assessment is an essential tool by which teachers have to influence the ways students respond to courses. On the other hand, there are clear drives from the UK government towards coursework-based assessment, focused on employability, that should apply across all degree subjects (Dearing, 1997; Browne, 2010). However, other studies show that not all assessment methods suit every programme and even courses (Gibbs, 1999). Thus, student assessment methods in HE can generally be divided into two main categories:

1- Exam-based assessment: which includes different forms such as closed and open book examination, essay-type exams and multiple choice exams.
2- Coursework-based assessments: which include research projects, assignments, reports, and presentations.

During the last few decades, the use of coursework-based module assessment has increased in the UK and other countries due to various educational arguments to justify its importance. Additionally, it appears to be that students prefer the assessment to be based on either coursework alone or a mix of both coursework and exams because these types of assessments tend to yield higher

marks than just exam-based assessment (Richardson, 2015).

The increased adoption of coursework-based assessment has contributed to an increase over time in the marks on individual modules, and in the percentage of good degrees across entire programmes (Bridges et al., 2002). Accurate and fair student assessment is an issue of concern in HE. Changes in the use of different assessment methods has given rise to the increasing number of universities that are shifting from traditional exam-based assessment to continuous assessment (i.e. coursework-based), (Heywood, 2000). Coursework-based assessment methods differ from exam-based assessment methods, where knowledge or skill is tested for a very specific period of time. Moreover, it has been widely acknowledged that the chosen assessment method will determine the style and content of student learning and skill acquisition (Heywood, 2000). Coursework marks are a better predictor of long-term learning of course content, compared to exams. (Graham Gibbs and Claire Simpson, 2005).

Different studies have proven that students tend to gain higher marks from coursework-based assignments, than they do from examinations (Morris and Fritz, 2015). Studies also found that combining exam-based and coursework-based assessment, as one assessment method, produced better average marks than did examinations alone (up to 12% higher average marks).

It appears to be that none of the studies have explored the relationship between the assessment method used in modules and students' final marks. Thus, one of the aims of this study is to isolate a factor/s from the assessment

method and use it again to re-evaluate student marks.

2.6 Conclusion:

Different studies that discussed, highlighted, and investigated the use of DM techniques in education, have been reviewed. Different research studies that employed EDM in analyzing higher educational data in particular have been reviewed.

The history of EDM, starting with the definition and applications, to the use of it in predicting student performance in HE, has been presented in this chapter, in addition to the methods and techniques that are used by researchers to draw conclusions regarding the relationship between initial data and predicted student performance. *Appendix C* of this thesis presents a summary of related work, the purpose of this summary is to show the most popular and important related work in term of students' academic performance in Higher Education, the categorizations have shown four important issues, first is presenting the objectives of the studies, second is focusing on the used data sources, third is showing the used data mining techniques and finally the results of these studies have been presented in detail.

EDM was defined through its key concepts, lifecycle, and applications, with an emphasis on predicting student performance and is a focus point of this thesis. Further reviews of studies that applied EDM to the analysis of the relationship between different factors of students' academic profiles, have been taken into consideration throughout the chapter. In addition, some DM

techniques that used to predict student performance in HE were reviewed and discussed.

Some arguments were put forward in this chapter, for example, regarding previous EDM studies which have applied the data preparation phase without considering loop backs within this phase. However, one of the main objectives of this thesis is to show how data can be repeatedly fed back into the dataset after applying different formulas during the data preparation phase.

All of the literature reviewed in this chapter use data collected from students' academic records as it stands and do not use formulas that represent more accurate and credible forms of the data. In this thesis, student attendance will not be analyzed in the conventional way where its value represents the attendance percentage/absence occurrences only, but also consider other factors, such as the number of times instructors recorded attendance. Moreover, a new method of measuring student attendance more accurately will be proposed in this thesis.

Another argument was put forward regarding the use of student marks as an attribute for predicting student performance. However, it appears that most of the previous studies have considered student marks as they are, neglecting the impact of the changes on student assessment methods. In fact, student marks should not be studied in isolation from the module's assessment method. This means that student marks should be prepared, during the data preparation phase, in a way that allows refining the marks, based on the assessment methods of the module.

Therefore, this thesis investigates the changes to recent assessment methods and tries to reflect these changes in the input variables, and student marks in particular. In chapter 2, the fourth objective has been achieved successfully by investigating the existing studies regarding students' academic performance, including the research objectives, data source and Data Mining (DM) techniques which have been applied.

Chapter 3 will describe the formulation of the research methodology adopted to achieve the objectives of this thesis. After examining the data and determining the objectives of the study, data cleaning methods, techniques, and tools to be used in this research will be outlined.

Data used in the research will be described in terms of source and type of data, and the research method used for such data, including all processes applied to prepare the data. The design of the experiments carried out in this research will be explained, including the DM techniques used and the reasons behind choosing them. Limitations and restrictions will also be detailed.

CHAPTER THREE: RESEARCH METHODOLOGY

3.1 Introduction:

In this thesis, the methodology used throughout the research will be described in detail in this chapter and coming ones. Processes represented by CRoss-Industry Standard Process (CRISP) for DM through which data went will be explained, from data understanding, to data preparation, the modelling and evaluation, and finally the deployment stage. In this thesis, deployment stage will be neglected as this study focuses only on theoretically applying the CRISP rather than being concerned with its business outcomes.

This chapter describes the preparation phase of the research methodology (CRISP-DM) adopted to achieve the objectives of this thesis. After examining the data and determining the objectives of the study, data cleaning methods, techniques, and tools to be used in this research, are specified.

In this chapter, data used in the research is described in terms of source and type of data, and research method that was used with such data. Additionally, processes that were applied to the data to prepare it, are also defined. Moreover, the design of the experiments undertaken is explained, with limitations and restrictions also detailed.

This research was intended to answer the following questions:

- Can student attendance and transcripts data have an alternative and more descriptive form than the ones currently used?

- Does the student attendance and transcripts data have a direct impact on students' academic performance?

3.2. Data Understanding:

In this section, data used throughout the research is described, including the source of the data, its type, and the processes that have been carried out to achieve clean and ready-to-use datasets.

3.2.1. The Data:

Data used in this research consists of two main parts: Student Attendance data and Student Transcript data. While the former represents module-oriented data, which reflects the characteristics of student attendance at each module, the latter characterizes students' transcript data with regards to the enrolled modules' assessment methods, students' marks at each level of their study, their module marks, and other transcript-related attributes.

A- Student Attendance Data:

Student attendance in Higher Education is an important issue as it has been shown to directly affect students' performance (Brijesh and Saurabh, 2011).

The student attendance data in this study represents data from 59 modules that was gathered from Computer Science undergraduates of a UK university during the years 2010-2015. This department was chosen due to the distribution of module assessment ratios, that will be explained in the following subsection.

Therefore, the main objective of handling student attendance data is to represent this data in a more formal and accurate form. This form can give an indication of student attendance at a certain module, alongside other factors that may have an effect on students' attendance, such as the lecturer's role in recording attendance.

B- Student Transcript Data:

The main objective of handling student transcript data is not only to highlight the effect of module assessment methods on students' academic performance, but also to formulate the different types of module assessment methods into a more accurate form.

Different studies have proved that students tend to gain higher marks from coursework-based assignments than they do from examinations (Morris and Fritz, 2015). Studies also found that combined exam-based and coursework-based assessment methods produced better average marks than did examinations alone (up to 12% higher average marks). Gibbs & Lucas (1987) reported an analysis of marks on more than 1,700 modules at Oxford Polytechnic. Modules with 100% coursework had an average mark which was 3.5% higher than modules with 100% examinations, and there were three times as many failed students on modules where there were only examinations.

The data investigated represents around 230,823 student records representing a total of six departments at a UK University, with each one of these department data sheets containing a number of student records. For each

record, several attributes that represent a student's academic accomplishments at three years (1st, 2nd and 3rd years) are divided as follows:

1) Student-Related Attributes: These attributes highlight the status of the students, including:

- Module Mark: a student's mark in a certain module.
- Exam Mark: the mark achieved by a student on the exam-based assessment.
- Coursework Mark: the mark achieved by a student on the coursework-based assessment.

2) Module-Related Attributes: A group of attributes that describe a certain module and its characteristics. These attributes include:

- Coursework Weighting (CWW): This attribute indicates the ratio of coursework-based assessment to the total mark of a module and can range from 100 (for a module assessed by coursework alone) to 0 (for a module assessed purely by examination).

This study confirms that students who are assessed using coursework only tend to get higher marks than those who are assessed using exams or a mixture of both coursework and exams, as shown in Table 3.1:

Table 3.1: Average module mark of all students in each department based on assessment method

Department	Number of Students	Average module mark of students		
		Exam-Based Assessment	Coursework-Based Assessment	Both Exam and Coursework-Based Assessment
Business	54960	59.77	60.83	60.01
Civil Engineering	34892	58.78	63.74	60.70
Computer Science	19800	58.18	64.40	58.87
Electronic and Computer Systems Engineering	13740	59.55	63.26	57.00
Math	24152	61.59	66.00	61.17
Mechanical Engineering	31385	58.80	64.26	60.24

A simple T-Test is applied to the data in Table 3.1 in order to measure the difference between the mean of each pair of variables. Results show that there is a statistically significant difference between the exam-based and coursework-based assessments (with a 95% confidence level; which equates to declaring statistical significance at the $p < 0.05$ level, a T-Value of -4.5 and a P-Value of 0.002).

Applying the T-Test to measure the significant difference between each pair of variables Ex-CW, Ex-Mix, and CW-Mix assessment methods results in Table 3.2:

Table 3.2. P-Values of T-Test for the variables Ex, CW and Mix assessment methods

Assessment Method	P-value	T-value
Exam and Coursework	0.002	-4.5
Mixed Assessment and Exam	0.749	0.39
Mixed Assessment and Coursework	0.004	-3.99

As shown in Table 3.2, the P-Value of the T-Test between the variables Exam and Coursework (0.002), mixed Assessment and Coursework (0.004), is less than 0.05, which indicates a statistical significant difference between these assessment methods. On the other hand, there seems to be no statistical significant difference between the Exam-Based Assessment and the mixture of both exam and coursework assessment methods, since the P-Value was 0.749, which is greater than the 0.05 level.

Throughout the literature, module assessment methods in HE have been investigated. It was shown in various studies that students tend to get higher marks when assessed using coursework than when assessed using exam-based assessment.

Table 3.1 shows that this study does not contradict existing studies, by confirming that students who are assessed using coursework tend to get higher marks than those who are assessed using exams or a mixture of both coursework and exams.

3.2.2. Data Preparation:

The data preparation phase of Educational Data Mining (EDM) includes all activities required to achieve the final data that will be fed into the modelling tools from the raw data. In addition, data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modelling tools (Wirth & Hipp, 2000).

1) Data Cleaning:

Educational Data is not normally measured using conventional methods as errors can be caused by humans (i.e. through data entry) so this data has minimum or zero noise, compared with other non-educational data types. However, in many cases, an instructor may intentionally or unintentionally skip recording a certain value in the data, which may result in a missing value (error). In the cases of either recording marks or attendance, these missing values are hard to fill, and since the percentage of the records with missing values, when compared to all data, in some cases is very small (0.000047% in the case of the Mechanical Engineering Department), those records were

neglected. Other filtering, sorting, and labelling processes were done to improve the visualization and readability of data. When the data became easily readable, the most significant attributes were chosen, based on their use; in other words, different attributes were chosen from the student attendance data to help with the formulation of student attendance, while other attributes were chosen from student transcript data to help with the formulation of module assessment methods.

2) Attribute Selection:

a) Student Attendance Data's Attribute Selection:

In this step, only the attributes that are required for processing with DM were selected from the original data (Appendix E) . The selection was based on the value those attributes add. In other words, some attributes add no value to the measurement per se, such as the module date or attendance reason. Therefore, these attributes were neglected, since it is not crucial to determine the average student attendance based on individual reasons as all occurrences count. However, other attributes represent the base of the measurement, such as the attendance status. All related variables selected from the student attendance data are given in Table 3.3.

The choice of attributes was based on the J48 Decision Tree's information gain. The J48 makes decisions by using the attributes with the highest normalized information gain and the algorithm recurs on the smaller subsets (Ross, 1992).

Table 3.3. Selected Attributes and their Information Gain

Attribute	Information Gain
Attend_Avg	2.357
Attend_Taken	0.728

b) Student Transcript Data's Attribute Selection:

Pearson Correlation was used to select the most significant attributes regarding student transcript data. Pearson Correlation coefficient is a statistical measure of the strength of a linear relationship between CW and EX marks. Table 3.4 shows the correlation coefficient of selected attributes where these values tend to be high for CW (0.69) and EX (0.966). These values indicate a strong linear relationship between CW to Module Mark and EX to Module Mark.

Table 3.4. Pearson Correlation of Data Attributes

	CW Mark	Exam Mark
CW Mark		
EX Mark	0.461	
Module Mark	0.690	0.966

By observing the correlation matrix above, it can be observed that the Exam Mark has the biggest positive affect on the Module Mark, followed directly by

the CW Mark. Therefore, this research has considered the most significant attributes, whose correlation values are the highest, in order to investigate the accuracy of students' module marks under different assessment methods.

3.3. Data Transformation:

As highlighted in the literature review, most of the studies that intended to predict students' academic performance, based on student attendance, student marks, or both, have been analysing the data with no transformation of this data. In other words, studies that considered students' attendance as a criterion for predicting their academic performance, have emphasized the attendance itself, with no focus on other factors that may affect the process of recording the attendance. For example, Yadav et al. (2012) has used student attendance as a criterion for predicting students' academic performance but other researchers have used this attribute on the form of an ordinal variable, whereby they considered students' attendance as either poor, average, or good. Moreover, Garey (1981) and David (1993) adopted student attendance rates (represented by the number of classes missed) as their criterion, to measure its affect on students' GPA.

Similarly, studies that adopted students' marks as criterion for predicting their academic performance, have isolated students' marks from any other attributes that affect those marks. In other words, whether adopting student marks as single module mark (Gedeon and Turner, 1993), or GPA (Osmanbegović and Suljić, 2012), the students' enrolled module assessment method should be

included as a variable in the prediction process since the assessment method (coursework-based or exam-based) affects a students' final mark.

Consequently, one of the main objectives of this thesis is to transform the student transcript and attendance data into new forms that help to link this data to other influential attributes. Specifically, student attendance data was formulated into Student Attendance Credibility (SAC) which links student attendance to instructors' frequency of recording attendance. Therefore, the attribute, Coursework Assessment Ratio (CAR) was used in order to take the different modules' assessment methods into account while preparing student transcript data, along with its effect on students' final mark.

In this thesis, the pre-processing part of Educational Data Mining (EDM) will be modified to improve the quality of the model outcome. The modifications of the pre-processing part of EDM will try to prove better outcomes using new proposed parameter engineering techniques such as SAC and CAR.

3.3.1. Student Attendance Credibility (SAC):

Credibility is form of statistical inference that uses newly observed past events to more accurately reforecast uncertain future events (Behan and Donald, 2009). In other words, it is a combination of experience estimates from historical data as well as base estimates in order to develop formulas. The formulas are used to replicate past experiences, and are then tested against actual data. In this thesis, the credibility of student attendance, which is a

measure of how much we believe and trust in students' attendance data that we used, will be investigated and formulated.

After preparing the data, the next step was to investigate which parameters may play a role in formulating the credibility of student attendance. The result of studying the most effective parameters is a proposed formula that can be applied to student attendance data, to achieve credible results, which in turn may help in evaluating the modules in terms of student attendance by measuring the credibility of each module and comparing the results.

Measuring SAC in HE can give an indication of the credibility of students' attendance, and can help in:

- Achieving a credible data source for different DM techniques.

- Evaluating the modules by accurately determining student attendance and absence, not by their number of occurrences, but rather depending on the effective weights of attendance and absence.

- Evaluating the credible student attendance for each module and comparing it to the same student's records in different years.

- Evaluating the performance of students based on their credible attendance or absence in each module, and then for all years of study.

There are many forms of students' attendance that are used in HE institutes such as, office hours, supervisors' meetings and workshops. However, this

study concentrates on student attendance in modules (classes) because it reflects the attendance or absence of a student in the module itself during the semester.

The main objective of this thesis is to measure the credibility of each module depending on two main factors: the average student attendance and the number of attendance occurrences taken by the instructor during the semester. A credibility equation has been formulated to measure the credibility of student attendance, which will be described step by step in the following section.

A. SAC Equation:

This section describes and validates the equation of SAC. This equation helps us to measure the credibility of student attendance; it can tell whether the module has a good attendance rate for a specific semester, depending on the following parameters:

- Average student attendance for each module: it is a significant parameter that represents the ratio between the number of students who attended in a certain week in each module, to the number of registered students in each module. Hence, it is a variable parameter that relies on both numbers (number of students who attended and number of overall registered students in the module).

- Number of attendance occurrences taken by the instructor for each module per semester. This significant parameter represents the number of attendance

occurrences taken by the instructor, and its significance comes from the fact that no information of student attendance will be available, whether the students really attended the module or not.

- Total number of weeks per semester: a variable parameter, depending on the system of each school, which should be part of the equation since it is connected directly to both the number of attendance occurrences taken by the instructor, and the average student attendance for each module.

1) Average Student Attendance

As mentioned earlier, the average student attendance depends on two factors: number of students attending in a certain week and the number of registered students on each module. Number of students registered on a module is given, hence let us assume that:

X= Number of students attending in a certain week on a module.

Y= Number of registered students on a module.

- Number of weeks, starting with 1 until the last week (n).

By dividing X over Y, we can obtain the ratio of attendance for one week only. Therefore, by considering the summation of this value throughout the semester multiplied by 100%, we can calculate the average percentage of student attendance for the whole semester.

To find the average student attendance, we have to use the following equation:

$$\text{Average Student Attendance} = \sum_{week=1}^{n} \left({}^{X}/_{Y}\right) * 100\% \qquad (1)$$

→ Note: the number of registered students in a module must be more than zero (otherwise there would be no module), therefore:

→ $Y \neq 0$

→ Note: Number of students attending in a certain week in such a module must be less or equal to the number of registered students in such a module, therefore:

→ $X \leq Y$

→ Average $\in$ [0-100%]

2) Credibility of Student Attendance

To measure SAC for each module, three parameters should be considered: the average value of student attendance should be calculated for each module; the number of attendance occurrences taken by the instructor for each module; and, finally, the total number of weeks for each semester. Let us assume that:

SAC: Student Attendance Credibility for each module.

A: Average student attendance for each module.

C1: Number of attendance occurrences taken by the instructor for each module.

C2: Total number of weeks for each semester.

Using the above notations, equation 2 can be derived, which measures SAC based on A, C1, and C2. The result will always be a number between 0 and 1. The purpose of finding SAC of each module is to find which module has a good attendance rate, compared with other modules. The values of the average student attendance for each module, the number of attendance occurrences taken by the instructor for each module and the total number of weeks for each semester, all affect the SAC value.

$$SAC = \frac{1}{100}\left(A * \frac{c1}{c2}\right) \qquad (2)$$

We can simplify the previous equation to be:

$$SAC = \frac{A\ c1}{100\ c2}$$

$$\frac{A}{100} * \frac{c1}{c2}\ , A \in [0\text{-}100]$$

Let us assume that Z1 denotes the attendance average divided by 100 (the left-hand side of the multiplication).

$$Z1 = \frac{A}{100}$$

$$Z1 \in [0\text{-}1]$$

Let us also assume that Z2, which is a ratio representing the instructor's overall recording of student attendance throughout the semester, denotes the number of attendance occurrences taken by the instructor for each module, divided by the total number of weeks for each semester (right- hand side of the multiplication).

$$Z2 = \frac{c1}{c2}$$

$$Z2 \in [0\text{-}1]$$

Therefore, the more attendance occurrences taken by the instructor, the closer to 1 Z2 will be.

$$\leftrightarrow \ 0 \leq Z1 \ \& \ Z2 \leq 1$$

The closer to 1 Z1 and Z2 are, the more reliable the student attendance record is (i.e. R approaches 1).

The limits of the credibility equation can be summarized as:

- when the number of weeks approaches n, equation (3).
- when the number of attendance occurrences taken by the instructor for each week approaches the number of weeks per semester (i.e. if the instructor is taking the attendance each time throughout the semester) and,
- at the same time, when the average value of student attendance is approaching 100% (i.e. when the number of students attending a certain module is equal to the number of registered students on the same module)

as in equation (4).

$$(\text{Lim } c2 \rightarrow \infty \frac{A\ c1}{100\ c2} = 0) \qquad (3)$$

$$(\text{Lim } c1 \rightarrow c2\ \&\ A \rightarrow 100 \frac{A\ c1}{100\ c2} = 1) \qquad (4)$$

From equations 3 and 4, it is clear that when the number of weeks increases, while fixing the values of A, and C1, the limit of SAC will approach 0, which means that in order to achieve higher credibility (i.e. =1) C1 and C2 must be equal (i.e. instructor must consider taking attendance every week during the semester), and A must be equal to 100 (i.e. number of students attending every week is equal to the number of students registered on the module).

As an example, equation 2 was applied to 59 modules and three different Computer Science modules (subjects) have been chosen and given the codes, X, Y, and Z. These represent a sample from the 59 modules' data that was gathered from Computer Science undergraduates at a UK university during the years 2014-2015. It was found that the first module (in this case, X) had a SAC ratio of 0.067, the second module (Y) had a SAC value of 0.349, and the third module (Z) had a SAC value of 0.812. In conclusion, the first module had the lowest credibility because its value was the smallest, which could be explained by a number of different factors, such as low number of attendance occurrences taken by the instructor for this module, or a low number of students attending the module in a certain week, compared to the total number of students registered on this module (i.e. low average). The third module had

the highest credibility, while the second module had medium credibility.

To distinguish between SAC and the occurrence of students' attendance on a certain module, values of SAC on each module should be compared. For example, in module X, the attendance of the student was 100%, knowing that the instructor took student attendance only once during the whole semester. Therefore, the value of 100% will not reflect the real attendance of students, since the module's SAC is very low (0.067). In the second module (Y), the attendance of the student was 50%, knowing that the instructor took the attendance seven times, which fails to reflect real attendance because the module's SAC is 0.349. On the other hand, for the third module (Z), student attendance was 70%, knowing that the instructor took student attendance eleven times (the total number of weeks in this semester). The 70% SAC ratio now reflects real attendance because the module's SAC is high (0.812 and Z2=1). Therefore, the higher the SAC, the more accurate the student attendance value will be.

Based on the above, SAC can be seen as an equation that not only takes into account the students' role as an indicator of their attendance, but also the instructor's role. The student's role is represented by A, since it measures student average attendance throughout the semester, while the instructor's role is represented by Z2 which refers to the ratio of taking attendance for the whole semester. That is, a perfect A does not reflect the student's real attendance, unless it is complemented with a perfect Z2, and vice versa (Alsuwaiket, et al., 2016). Table 3.5 below summarize the example for all

cases:

Table 3.5. Credibility of student attendance example

Module Code	SAC Ratio	A	C1	C2	Z2
X	0.067	100%	1	11	0.09
Y	0.349	50%	7	11	0.63
Z	0.812	70%	11	11	1

In order to consider the new measure (SAC) for future uses, its consistency should be assessed. For this purpose, SAC is measured throughout a period of five academic semesters from 2010 to 2015 and SAC values are then compared so that Cronbach's Alpha coefficient can be measured to determine SAC's consistency (Mohsen T. and Reg D., 2011).

The way that usually used to measure the consistency is Cronbach's alpha, a statistic calculated from the pair-wise correlations between items. In addition, the ranges of consistency are between $-\infty$ and 1. Coefficient alpha will be negative whenever there is greater within subject variability than between subject variability (Knapp, 1991).

A commonly accepted rule of thumb for describing internal consistency is as follows in Table 3.6 (George & Mallery, 2003).

Table 3.6. Commonly accepted rule of thumb for describing internal consistency

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

Ten modules taken by students every year have been chosen randomly, amongst other modules. SAC was measured for the modules for each year, starting from the academic year 2010/2011 until 2014/2015. The results are shown in Table 3.7.

With the aim of obtaining a measurement of consistency, Cronbach's Alpha internal consistency was applied; though it defines the consistency of the results delivered in a test, ensuring that the various items measuring the different constructs, deliver consistent scores (Reynaldo and Santos, 1999). However, in this study, it was used to calculate the consistency of the SAC measure for 10 modules, over a period of 5 years. In fact, the random sample of 10 modules has taken because of the consistency equation takes random sample during period of time. For example, for the first module

(XXCOA101), the SAC value 0.596 has been calculated for the year 2010/2011, the same has been applied for all years between 2010 and 2015 for this module. On other hand, the SAC values have been calculated for the remaining modules to be used for further calculations to Measure the value of Cronbach's Alpha.

Table 3.7 shows the processes applied to the data, resulting in an Alpha value of 0.81, which ensures high consistency of SAC.

Table 3.7. SAC for 10 Random Modules Over 5 Years and Measuring Cronbach's Alpha of SAC

	Year10/11	Year11/12	Year12/13	Year13/14	Year14/15	TOTAL
XXCOA101	0.596	0.585	0.641	0.720	0.801	3.343
XXCOA122	0.543	0.681	0.636	0.574	0.552	2.986
XXCOB101	0.412	0.407	0.439	0.557	0.601	2.416
XXCOB231	0.223	0.482	0.399	0.265	0.538	1.906
XXCOB232	0.335	0.389	0.417	0.416	0.505	2.063
XXCOB290	0.340	0.564	0.496	0.623	0.713	2.735
XXCOB301	0.559	0.466	0.628	0.358	0.592	2.604
XXCOC003	0.389	0.423	0.371	0.444	0.556	2.183
XXCOC101	0.538	0.505	0.536	0.593	0.703	2.875
XXCOC104	0.484	0.464	0.503	0.428	0.489	2.368
Total	4.419	4.966	5.067	4.978	6.050	25.481
Var	0.015	0.008	0.010	0.019	0.010	0.063

k	5
∑Var	0.063
var	0.180
α	0.815

Where:

k: is the number of years.

Var: is the variance of SAC values for each module over one year only.

∑Var: is the sum of Var values over a period of 5 years.

var: is the population variance.

α: is the Cronbach's Alpha, whose values can be between 0 and 1.0 (the higher, the more consistence) and its equation can be given as:

$\left(\frac{k}{k-1}\right) * \left(1 - \frac{\sum var}{var}\right)$ …Cronbach's Alpha (Lee J. Cronbach, 1951).

The equation of Cronbach's used the parameters shown in Table 3.7 to calculate the alpha value, Based on the above results of the processes applied to the data, resulting in an Alpha value of 0.81, which ensures high consistency of SAC which represents the consistency of the SAC measure for 10 modules, over a period of 5 years. It can be concluded that the newly proposed measure (SAC), is a good and reliable measure when dealing with student attendance data; hence it can be considered as a consistent measure.

3.3.2. Coursework Assessment Ratio (CAR):

To refine students' marks based on module assessment methods, the data has been processed at various stages and each of these stages enhances the data in terms of the readability by statistical software and the ability to easily extract information. The first step is to categorize the CW to EX ratios. The categorization algorithm relies on the number of classes of the ratio between CW to EX Weighting can have, based on ratios of CW to EX the original data have. Namely, each department has its own classification of CW to EX Weightings ratios.

CAR is a coursework assessment ratio that represents the ratio of CW weighting for each module, which by default will reflect the ratio of EX weighting. It is necessary to categorize the ratio values in Table 3.8 because it is statistically difficult using statistical software. Table 3.7 shows the different classes:

Table 3.8: Classes of CW to EX Weighting Ratios

	Model Assessment Method																
CW	0	10	15	20	25	30	35	40	45	50	55	60	65	66	70	75	100
Business	✓	✓		✓	✓	✓		✓		✓		✓	✓				✓
CEng	✓		✓	✓	✓	✓		✓	✓	✓						✓	✓
CS	✓	✓		✓	✓	✓		✓		✓	✓	✓	✓		✓		✓
ECSEng	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓		✓			✓
Math	✓	✓		✓	✓	✓		✓		✓					✓		✓
MEng	✓	✓		✓	✓	✓		✓		✓					✓	✓	✓

Table 3.8 shows that there is no single department which shares the same ratio classes with other departments, i.e. each department has its own unique ratios between CW to EX. Thus, filling the missing values in the table is not a solution, since doing so yields incorrect data. This research considers the department with the most number of classes to start with, and then generalizes the findings to other departments, while bearing in mind the change of ratios. From the two departments that have 12 complete ratios (the CS and ECSEng departments), the CS department was chosen for further processing.

A) Refining Student Marks based on CAR

In order to uncover the relationship between the student's module mark and CAR, which represents the CWW to EXW ratios as mentioned before, simple quadratic regression is used. Regression analysis is being used to infer relationships between the independent (CAR) and dependent (Module Mark) variables. The variable of CW Ratio was used as simple quadratic regression is more suitable for one variable relation. The choice of choosing quadratic over linear is based on the R-squared; when the R-squared is higher, the better the model fits the data. It is also known as the coefficient of determination of the model. For the case of quadratic regression, the coefficient is 2.90%, which is higher than it is in the linear model (2.77%). By applying simple quadratic regression to the data, with CAR as a variable for the module mark as a response, we achieved the following fitted regression line:

$$RMM = MM - 12.77\,(CAR) + 5.873\,(CAR)^2 \qquad (5)$$

Where RMM is the refined module mark after fitting and MM is the current module mark.

By applying the above equation on the student transcript data of the CS Department, an additional field will be added which contains the RMM for each student at the department.

RMM are self-explained when referring to Table 3.1, that compares the average module marks for students attending modules, according to different

assessment methods. That is, the higher the percentage of CWW, the more the added marks to MM, and vice versa.

Any EDM studies that consider student marks with different assessment methods should consider adding a sub-process within the data pre-processing phase, that takes into account the difference between those assessment methods.

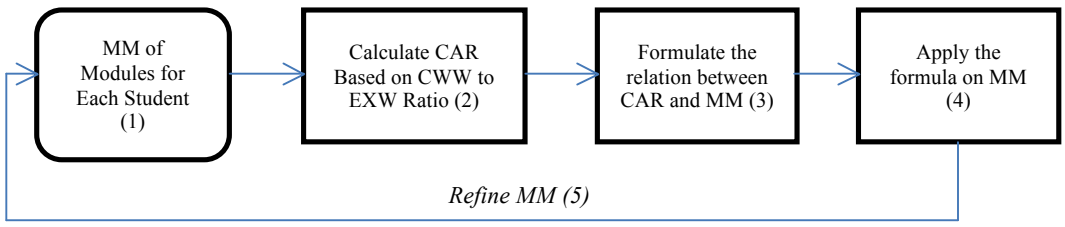


Figure 3.1. Refining Students' Module Marks Sub-Process

Figure 3.1 shows the addition sub-processes 2-5 on raw students' module marks. The task of theses sub-processes is to take the differences in assessment methods into account. This sub-process also takes into consideration the fact that applying conventional DM processes on educational data may not produce accurate results.

This study shows that the pre-processing phase of educational data should include additional sub-phases that deal, not only with noise or missing data, but also with data refinement to cope with differences between various educational systems and their data. Therefore, an attribute (CAR) was

constructed in order to take the different modules' assessment methods into account while preparing student transcript data.

By refining students' marks, they either increase or decrease depending on the relationship between CWW and EXW, for each student during their study. For instance, by considering student X as an example, the student had enrolled in 32 modules during his/her study at the Department of CS. Nineteen out of thirty-two modules are 100% exam-based assessed modules; seven modules are assessed by a mixture of coursework and examination, while 6 modules are 100% assessed by coursework only. So, the majority of the modules are assessed through examination only which means that the student gets no extra marks, compared with coursework-based modules and the other modules, which allow students to gain extra marks and hence add to their overall average. In numbers, nineteen modules have a 0 CAR value, which means that RMM = MM. On the other hand, the rest of the modules have values of CAR from 0.1 to 1, which means that the RMM is always less than MM for those thirteen courses. This decrement in the marks is due to the fact that students get higher marks in modules that are assessed by coursework or a mixture of coursework and examinations. Therefore, to balance the module marks and the overall average, the formula decreases the module marks by varying percentage, depending on the CWW to EXW ratios. Table 3.9 shows the differences in module marks and overall average for the example student X.

Table 3.9. Refining student "*X*" Marks Based on Enrolled Modules' Assessment methods

	Average MM	**Average RMM**
19 Exam-Based Modules	48.6	48.6
6 Coursework-Based Modules	60.3	52.7
7 Mix of EX and CW Modules	60.4	58.3
Total of 32 Modules	56.4	53.2

By following the procedure in Figure 3.1 on a real student's marks, Table 3.8 shows that RMM remain unchanged for the student, when the assessment method of the enrolled modules is purely exam-based. Alternatively, when the assessment method of the enrolled modules is purely coursework-based, the RMM is reduced by an average 7.6 marks, compared to MM. Finally, a mixture of EX and CW based modules yields less refinement of the RMM (2.1 marks) compared to MM for the same student. This means that student X in the example, who is taking 32 modules of different types, may have his average marks reduced by 3.2%, when applying the proposed procedure.

SAC and CAR are two constructed attributes that link student attendance to instructor's frequency of recording attendance, and re-evaluate student marks based on the effect of module assessment methods on those marks, respectively. These attributes will be considered for predicting students' academic performance.

3.4 Conclusions:

In Chapter 3, the methodology used in this thesis has been described. First, the source of used data and its type were characterized. Data used in this research consists of two main parts: student attendance data and student transcript data. While the former represents module-oriented data, which reflects the characteristics of student attendance at each module, the latter characterizes students' transcript data regarding the enrolled module assessment methods, students' marks at each level of their study, their module marks, and other transcript-related attributes.

Then, the data preparation process was explained, through which data was cleaned, sorted, filtered, transformed a few times in order to suit different DM techniques and different phases of the experiments, and various new attributes were constructed to serve this purpose. Two main attributes were constructed and formulated to differentiate this research from previous research looking at students' academic performance. The newly formulated attributes: Student Attendance Credibility (SAC) and Coursework Assessment Ratio (CAR) have been outlined and discussed.

SAC tries to formulate student attendance in a way that records the number of times a student attends classes but also instructors' frequency of attendance recording. On the other hand, CAR tries to improve the method for analysing the link between students' enrolled module assessment method and the students' marks, since the literature highlighted varying degrees of impact that module assessment methods have on students' marks. However, it appears

that no studies have reflected this difference back on students' marks.

In chapter 3, the first three objectives have been achieved successfully by investigating the factors of students' academic profiles from the data set used, identifying those which have the biggest impact on their academic performance, using those factors in a way that guarantees more accuracy and credibility and showing the differences between using data collected from students' academic records as it is, and using formulas that represent these data in a more accurate and credible form. Therefore, student attendance will not be recorded using conventional methods, where its value represents number of attendance/absence occurrences only, but also considering other factors such as the number of times instructors record attendance. Similarly, student marks will be looked at in relation to the module's assessment method.

First, third and fourth contributions have been achieved successfully by Modifying the Educational Data Mining (EDM) pre-processing part to improve the quality of outcome, formulating students' transcript data in a way that guarantees the assessment methods of students' enrolled module reflects the original data. In other words, a contribution of this thesis is to consider the differences between various assessment methods and refine student marks based on these differences and Formulating student attendance in a way that takes into consideration other factors that have an impact on students' attendance (such as number of times the instructors record attendance), rather than using the attendance data as it is (e.g. number of absences, poor or good attendance, etc.).

CHAPTER FOUR: EXPERIMENT WORK

4.1 Introduction:

Chapter 3 the source of data and their types were characterized. Data used in this research consists of two main parts: student attendance data and student transcript data. The data preparation process was also explained. Two main attributes were constructed and formulated to differentiate this work from previous work that tackled the problem of students' academic performance. The newly formulated attributes: Student Attendance Credibility (SAC) and Coursework Assessment Ratio (CAR).

EDM process converts raw data from educational systems into useful information that could potentially have a greater impact on educational research and practice (Romero, et al., 2010). Additionally, EDM uses a wide range of methods to analyse data, including, but not limited to, the Supervised and Unsupervised model induction, parameter estimation, Relationship Mining, and other methods (Romero and Ventura, 2007; Baker and Yacef, 2009). This chapter describes in detail the steps followed to achieve various predictive models of students' academic performance. More specifically, the steps and results of the following three main DM techniques will be explained: Random Forest (RF), Artificial Neural Networks (ANN) and k-Nearest Neighbours (kNN).

Additionally, this chapter will discuss the data used in the research, the processing steps applied to the data, the required tools, and the results of applying DM tools on the modified data.

4.2 The Data:

Data that was used with DM techniques had been pre-processed through different stages. The stages included adding novel attributes that increased the accuracy of prediction (to be shown in this chapter) and other various pre-processing stages to guarantee clean and ready data.

Data used in this research represented either module-centred or student-centred data. In other words, the former represents all data relevant to a module, including its assessment method, the ratio of split between coursework and exam-based assessment, students enrolled in a module, their marks, their level (year) of study (A, B, and C), and other module-related attributes. This data lies under the Student Transcript Data's umbrella. On the other hand, the student-centred data reflects all attributes related to a student's attendance in modules, also known as Student Attendance Data. (Appendix F)

4.2.1 Calculating Student Averages for Each Level:

Student Transcript Data included different attributes that are related to student marks for various modules, level of students (A which represents first year, B for second year, C for third year, and other relevant levels), and other associated factors. The first step is to clean the data to isolate students with only three levels (A, B, and C), then calculate the average mark for each

student at each level, and then calculate the total average for each student for all levels. (Appendix D)

4.2.2 Coursework Assessment Ratio (CAR):

Coursework Assessment Ratio (CAR) was first calculated in a way that represents the ratios between coursework-based and exam-based modules. For instance, CAR may equal 0.25, which represents 25% coursework and a 75% exam-assessed module.

For each enrolled module, every student has a new attribute indicating the CAR for that student on the enrolled module. Additionally, a student enrolled on various modules with different CAR values has an average CAR, which represents the sum of CAR values, divided by the number of enrolled modules. For example, a student attending 5 modules with CAR values of (0.3, 0.6, 0.9, 0.2, and 0.75) will have an average CAR of 0.55.

4.2.3 Average Student Attendance Credibility (SAC):

Student Attendance Credibility is the measure by which educators can achieve more accurate recordings of student attendance (Alsuwaiket., et al., 2016).

In Alsuwaiket et al (2016), SAC was measured per module and that was useful for measuring the credibility of student attendance at modules. However, in this chapter, SAC will reflect each student in order to measure, not only the credibility of student attendance at a module, but also show how credible a

certain student's attendance was. The procedure followed to achieve this goal was as follows:

- Firstly, grouping student attendance data records into groups of students with the same enrolled modules.
- Secondly, for each student at each module, average attendance was calculated (number of attendance occurrences/total number of lectures for the module).
- Thirdly, the factor C1/C2 (which represents the number of times the instructor recorded attendance/total number of classes) was calculated for each student.
- Finally, SAC was calculated for each student based on the formula:

$$1/100 *(C1/C2 * A) \qquad (6)$$

Now, by taking a certain student as an example, the student may take different modules during each level of his study and in this case, Average SAC is the average of all SAC values for this student, during a certain level.

After preparing the data, the final form of it was achieved. Table 4.1 shows an example of the achieved final form of data:

Table 4.1: Example on the Final Form of Data

Student Number	Average A (Year1)	Average B (Year2)	Average C (Year 3)	Average SAC	Average CAR	Total Marks Average
Example 1	70.94	88.32	78.07	0.45	0.75	79.11
Example 2	55.50	62.55	59.61	0.65	0.55	59.22
Example 3	47.83	51.18	63.44	0.49	0.3	54.01
Example 4	63.78	41.83	54.2	0.34	0.25	45.19
Example 5	55.64	53.1	67.9	0.59	0.5	55.58
Example 6	70.78	69.4	80.2	0.85	0.25	71.97
Example 7	53.89	60.8	61.8	0.66	0.3	57.99
Example 8	79.56	74.4	79.1	0.73	0.25	73.92
Example 9	63.89	53.18	69.1	0.72	0.3	57.83
Example 10	52.38	61.09	66.9	0.78	0.25	61.17

4.2.4 Notes on the Final Form of Data:

Obviously, many students may study one or two years only, or they may continue their entire program and study all three years. This, in fact, was reflected in the final modified data, where the number of records for students of level A (788 records) is greater than both students of level B (606 records) and students of level C (406 records).

Except for the Student Registration Number which is a Meta attribute (string of alphanumeric digits), all other attributes are continues numerical attributes (real numbers), and this has led to transforming the data, the total average mark in particular, into two other types: Alphabetical Nominal, and Numerical Nominal attributes, to be shown in the following section.

Data records were split into 70:30 ratio records; the 70% represents the training data and the 30% is the testing data. The Total Marks Average attribute of the testing data was hidden, since it is the target of prediction.

4.3 Data Transformation:

To choose the DM techniques for predicting student performance, data had to be transformed multiple times in order for the suitable DM technique to be identified. That is, some DM techniques do not accept the output (target of prediction) to be continuous, thus it was necessary to transform this attribute into another form. Since some DM techniques require a nominal attribute in

order to perform its prediction task, the Total Averages attribute was transformed as follows:

Table 4.2 Continues Total Average to Nominal Equivalent

Total Averages	British Degree Classification (Loughborough University)
[70-100]	First
[60-70)	Upper Second
[50-60)	Lower Second
[42-50)	Third
[40-42)	Pass
Below 40	Fail

4.4 Orange Canvas (Data Mining Tool):

Orange Canvas is a Python-based tool for DM, Orange offers a structured view of supported functionalities grouped into the following categories: visualization, evaluation, regression, unsupervised learning, prototype implementations, classification, data operations, and association. Supported functionalities are visually represented by different widgets such as read file, discretize, train SVM classifier and others.

In some other tasks and DM techniques, Orange was the choice for its ease-of-use, visual programming, and interactive data visualization characteristics. In particular, Orange was used to:

- Compare the accuracy of different DM techniques.
- Apply DM techniques.
- Show and compare results of applied techniques

4.5 Choosing the Data Mining Techniques:

In the literature, it was shown that numerous DM techniques are being used to fit specific data for precise tasks (Wahbeh and Al-Radaideh, et al. ,2011). In this research, data has been fed into the Orange DM tool, and then a comparison was carried out between different DM techniques in order to decide which DM techniques have more prediction accuracy than others, and based on that a decision was made.

Data were fed into a prediction comparison model as shown in Figure 4.1. Attributes (features) for prediction were Average A, Average B, Average C, Average CAR, and Average SAC, where the total average was the target of prediction.

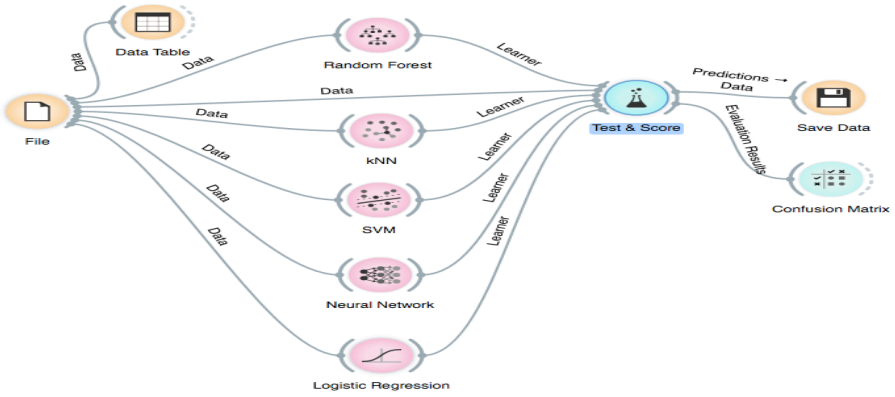


Figure 4.1. Orange Model for Comparing Different DM Techniques

In Figure 4.1, data was divided into two parts: one for the training and the other for testing data. The DM techniques used in this model are Random Forest, Neural Network, kNN, SVM and Logistic Regression. After applying the data, the prediction model gave the evaluation results for this model. Finally, the Confusion Matrix was generated to show how the data is classified, either correctly or incorrectly.

Table 4.3. Comparison of Prediction Accuracy between the Most Common Prediction DM Techniques

	Evaluation Results			
Method	**AUC**	**CA**	**MSE**	**RMSE**
Random Forest	1.000	0.929	0.044	0.209
Neural Network	0.998	0.894	0.152	0.389
kNN	1.000	0.892	0.101	0.317
SVM	1.000	0.889	0.175	0.418
Logistic Regression	0.963	0.554	0.446	0.667

As shown in Table 4.3, a comparison has highlighted the most accurate DM techniques, in terms of four standard measures, that predict students' total averages based on our data and attributes.

1. ***Area Under the Curve (AUC):***

AUC represents the proportion of false positive rate covered by the curve of true positive rate and is considered a measure of the classifier's performance

(Andrew, 1997). Moreover, the AUC value will always be between 0 and 1.0 because it is a portion of the area of the unit square. However, due to the random guessing produces the diagonal line between (0, 0) and (1, 1), which has an area of 0.5, the realistic classifier should have an AUC of larger than 0.5.

The AUC has an important statistical property; the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

The AUC value reflects the overall ranking performance of a classifier and is calculated as:

$$AUC = \frac{S_P - n_P(n_n+1)/2}{n_P n_n} \qquad (7)$$

Where S_P is the sum of the positive examples ranked, while $n_P n_n$ denotes the number of positive and negative examples respectively.

2. ***Classification Accuracy (CA):***

CA represents the accuracy of the classifier and refers to the ability of the classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of a predicted attribute for new data. CA can be measured by calculating the percentage of correct prediction, divided by the total number of predictions (Kotsiantis et. al, 2007).

Table 4.4. General Confusion Matrix

A\P	Predicted Yes	Predicted No	
Actual No	**TP**	**FN**	**P**
Actual Yes	**FP**	**TN**	**N**
	P'	**N'**	**All**

As shown in Table 4.4, the Confusion Matrix has some basic terms which described below:

- **True Positives (TP):** These are cases in which predicted yes.
- **True Negatives (TN):** These are cases in which predicted no.
- **False Positives (FP):** These are cases in which predicted yes, but they actually yes.
- **False Negatives (FN):** These are cases in which predicted no, but they actually yes.
- **P':** Total Number of Predicted Yes
- **N':** Total Number of Predicted No
- **P:** Total Number of Actual Yes
- **N:** Total Number of Actual No

The Confusion Matrix can generate the Classifier Accuracy (CA), or recognition rate: percentage of test set tuples that are correctly classified

Accuracy = (TP + TN)/All

Error rate: 1 – accuracy, or Error rate = (FP + FN)/All

3. *Mean Square Error (MSE):*

MSE is a measure of the quality of an estimator; it is always non-negative, and values closer to zero are better (Lehmann, 1998). In general, the MSE measures the difference between the predicted solutions and desired solutions. A smaller MSE value is required in order to gain better results. The MSE is defined as below:

$$MSE = \frac{1}{n}\sum_{1}^{n}\left(P_j - A_j\right)^2 \qquad (8)$$

"Where Pj is the predicted value of instance j, Aj is the real target value of instance j and n is the total number of instances. Through the learning process, the solution that has a minimum MSE score, will be used as the final model (best solution)" (Lehmann, 1998).

Similar to accuracy, the main limitation of MSE is this metric does not provide the trade-off information between class data. This may lead the discrimination process to select the sub-optimal solution. Moreover, this metric is highly dependent on the weight initialization process. In extremely imbalanced class problems, if the initial weights are not properly selected (i.e. no initial weight to represent the minority class data), this may lead the discrimination process to end up with a sub-optimal solution due to a lack of information of minority class data, even though the MSE value is minimized (under-fitting or over-fitting).

4. ***Root Mean Square Error (RMSE)***:

The RMSE is frequently used to measure the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed (Lehmann, 1998). On other hand, there is another term that would give the same meaning of RMSE called Root Mean Square Deviation (RMSD), this measure have frequently used of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These differences are called residuals, and the RMSE collecting them into a single measure of predictive power.

The RMSE of a model prediction, with respect to the estimated variable X model, is defined as the square root of the mean squared error (Lehmann, 1998):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs}-X_{model,i}\right)^2}{n}} \qquad (9)$$

Where $X_{obs}$is observed values and $X_{model}$is modelled values at time/place i.

The RMSE values can be used to distinguish model performance in a calibration period with that of a validation period, as well as to compare the individual model performance to that of other predictive models.

Random Forest has scored the highest CA value (0.929) and the lowest RMSE value (0.209) Neural Network had a CA value (0.894) and RMSE value (0.389), kNN had a CA value (0.892) and RMSE value (0.317), which

predicts output according to the nearest training instances (Imandoust, S. B., & Bolandraftar, M., 2013). Support Vector Machine (SVM) (Huang & Fang, 2013) scored a CA value (0.856) and RMSE value (0.418), and finally, Logistic Regression (LR) scored a CA value (0.554) and RMSE value (0.667).

It must be mentioned that this rank may differ if the same DM techniques were used on different data. Based on this rank, Random Forest (section 4.6.1), Neural Network (section 4.6.2) and kNN (section 4.6.3) were chosen for further processing and a more detailed prediction of students' academic performance.

In addition, Logistic Regression (LR) and Support Vector Machine (SVM) have some issues related to the nature and size of the data that may affect the accuracy of the results. In term of applying LR, one of its advantages that it can accept big number of independent variables, and there are many occasions when it is accepting small number as well. However, LR would work accurately when the size of sample is bigger better that dealing with small ones.

Unlike linear regression, logistic regression can only be used to predict discrete functions. Therefore, the dependent variable of logistic regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data. Here it can be concluded that using LR will not be useful due to the size and type of chosen data in this thesis.

SVM can be a useful technique for insolvency analysis, in the case of non-regularity in the data, for example when the data is not regularly distributed or has an unknown distribution. It can help evaluate information, i.e. financial ratios which should be transformed prior to entering the score of classical classification techniques. Here it can be concluded that using SVM will not be useful due to the type of chosen data in this thesis.

4.6 Applying Data Mining Techniques:

There have been several studies that tried to predict students' academic performance using many forms (final grade, Grade Point Average (GPA), or other indicators), for instance Bayesian networks have been used to predict student performance (Haddawy, et al., 2007) and to predict a future graduate's cumulative GPA based on students' background at the time of admission (Hien & Haddawy, 2007). On the other hand, similar research has been carried out on student marks, in order to predict their final grade (Gedeon & Turner, 1993). Similarly, other DM techniques were used for the same purpose: Decision Trees and association rules have been used to predict student performance (Nebot, et al., 2006) (Chan, 2007).

Nonetheless, it appears that none of the related studies have either incorporated students' enrolled module assessment methods or formulated student attendance in a way that may increase the accuracy of prediction.

The aim of applying DM techniques in this research is to predict second year average marks of students, given their first year's average, predict third year's

average mark of students given their first and second year's average and predict third year's average mark of students given their first year's average. Average SAC, and Average CAR will be used in order to highlight any improvements on the prediction accuracy.

- Input Data:

Data tested through the experimental work, transcript data in particular, contained end-of-year average, average mark, and calculated Students Attendance Credibility (SAC) and Coursework Assessment Ratio (CAR) for each student. Dataset have been divided into two samples, one for training (70%) and the other one for testing (30%).

Experiments of DM techniques used will investigate the following:

- Predict final students' marks of the second year based on their first year's marks.
- Predict final students' marks of the third year based on their first and second years' marks.
- Predict final students' marks of the third year based on their first years' marks only.

Additionally, in terms of SAC and CAR involvement in predicting students' mark, further investigation will be carried out. Some scenarios are as follows:

- Predicting students' average marks while excluding SAC and in another experiment including it and comparing the results.

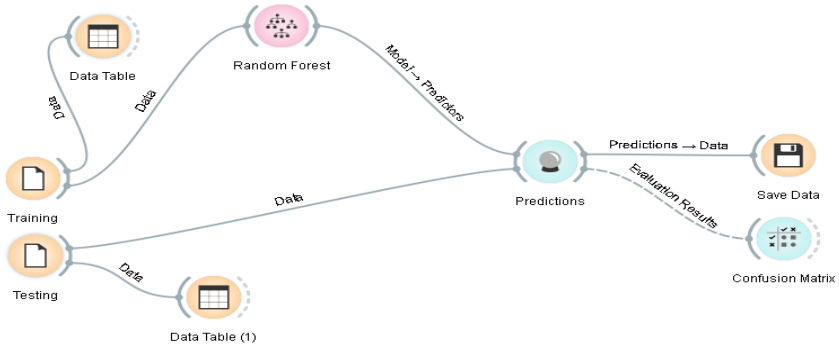
- Predicting students' average marks while excluding CAR and in another experiment including it and comparing the results.
- Predicting students' marks while including both CAR and SAC, firstly including the average mark of the first year only, then including the first and second years, and comparing the results.

4.6.1 Random Forest (RF):

Random Forest (RF), a classification method, is essentially a DM package, based fundamentally on regression tree analysis and feature importance (Breiman, 2001). The Random Forest DM technique has been used for predicting student performance (Kotsiantis, et al., 2010), (Yadav, et al., 2012), (Romero, et al., 2013).

Breiman (2001) states that "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest".

Figure 4.2 Applying Random Forest DM Technique to the Data



In Figure 4.2, data was divided into two parts: one for the training and the other for testing data. The DM technique used in this model was RF and after applying the data to RF, the prediction model gave the evaluation results for this model. Finally, the Confusion Matrix was generated to show how the data is classified, either correctly or incorrectly.

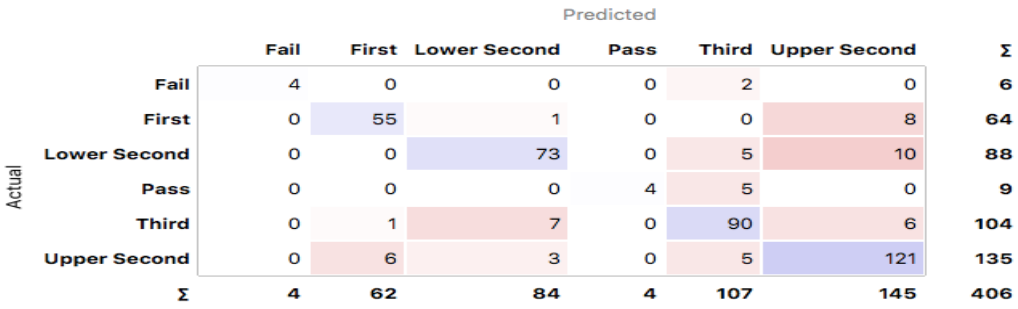
The Random Forest DM technique was applied to the data (as shown in Figure 4.2) using the Orange Canvas tool. The following four sections will represent the implementations of the Random Forest technique, in terms of using SAC and CAR in predicting students' marks. Each section has three different scenarios of data input, as mentioned in the previous section. In chapter five, all results will be interpreted and evaluated in detail.

4.6.1.1 Applying Random Forest Technique (With SAC & CAR):

In this section, the RF technique was applied to the data, including Student Attendance Credibility (SAC) and the Coursework Assessment Ratio (CAR). Three different scenarios were applied, firstly to predict B average marks from level A average marks, secondly to predict C average marks from level A and B average marks, and thirdly to predict C average marks from level A average marks.

The RF Confusion Matrix of Predicted B average marks from Level A average marks, is given in Table 4.5.

Table 4.5 RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks



Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	4	0	0	0	2	0	6
First	0	55	1	0	0	8	64
Lower Second	0	0	73	0	5	10	88
Pass	0	0	0	4	5	0	9
Third	0	1	7	0	90	6	104
Upper Second	0	6	3	0	5	121	135
Σ	4	62	84	4	107	145	406

Table 4.6 RF Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.988	0.141

The result of applying Random Forest, when including SAC and CAR to predict B Average marks from level A average marks, has a very high CA value of 0.988 and a very low RMSE value of 0.141. The Random Forest in this case starts by dividing the data into two groups, then again and again, depending on data values, until it reaches its target.

Table 4.7 RF Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Predicted

Actual	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	15	0	1	0	0	2	18
First	0	104	2	0	0	5	111
Lower Second	1	0	55	0	1	13	70
Pass	0	0	0	4	0	0	4
Third	2	0	3	0	28	5	38
Upper Second	0	5	1	0	0	159	165
Σ	18	109	62	4	29	184	406

Table 4.8 RF Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.973	0.184

The result of applying Random Forest when including SAC and CAR to predict C Average Marks from Level A and B Average Marks, has a very high CA with a value of 0.973 and a very low RMSE of 0.184 (as shown in Table 4.8).

Table 4.9 RF Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	14	0	0	0	2	2	18
First	0	100	2	0	0	9	111
Lower Second	1	2	55	0	0	12	70
Pass	1	0	1	2	0	0	4
Third	1	0	3	1	26	7	38
Upper Second	0	9	2	0	1	153	165
Σ	17	111	63	3	29	183	406

Table 4.10 RF Evaluation Results of Predicted C Average Marks from Level A Average Marks

	Evaluation Results	
Method	**CA**	**RMSE**
Random Forest	0.963	0.228

The result of applying Random Forest when including SAC and CAR to predict C Average Marks from Level A Average Marks, has a high CA value of 0.963 and a low RMSE of 0.228 (as shown in Table 4.10).

4.6.1.2 Applying Random Forest Technique (With SAC only):

In this section, the RF technique was applied to the data, including Student Attendance SAC only. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C

Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.11.

Table 4.11 RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Predicted

Actual	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	3	0	0	2	1	0	6
First	0	59	0	0	0	5	64
Lower Second	0	0	71	0	8	9	88
Pass	0	0	0	5	4	0	9
Third	0	1	6	1	87	9	104
Upper Second	0	4	7	0	3	121	135
Σ	3	64	84	8	103	144	406

Table 4.12 RF Evaluation Results of Predicted B Average Marks from Level A Average Marks

	Evaluation Results	
Method	**CA**	**RMSE**
Random Forest	0.975	0.207

The result of applying RF, including SAC only to predict B Average Marks from Level A Average Marks, has a high CA with a value of 0.975 and has a low RMSE of 0.207 (as shown in Table 4.12).

Table 4.13 RF Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	15	0	0	0	1	2	18
First	0	95	1	0	0	15	111
Lower Second	1	1	50	0	3	15	70
Pass	0	0	1	2	0	1	4
Third	1	0	3	0	27	7	38
Upper Second	0	8	3	0	1	153	165
Σ	17	104	58	2	32	193	406

Table 4.14 RF Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.961	0.204

The result of applying RF when including SAC only, to predict C Average Marks from Level A and B Average Marks, has a CA with value 0.961 and an RMSE of 0.204 (as shown in Table 4.14).

Table 4.15 RF Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	13	1	0	0	2	2	18
First	0	103	0	0	0	8	111
Lower Second	1	3	51	0	2	13	70
Pass	0	0	2	0	1	1	4
Third	0	0	1	0	26	11	38
Upper Second	0	10	5	0	0	150	165
Σ	14	117	59	0	31	185	406

Table 4.16 RF Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.958	0.221

The result of applying RF, when including SAC only, to predict C Average Marks from Level A Average Marks, has a CA with a value 0.958 and an RMSE of 0.221 (as shown in Table 4.16).

4.6.1.3 Applying RF Technique (With CAR only):

In this section, the RF technique was applied to the data, including CAR only. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level

A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.17.

Table 4.17 RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	3	0	0	0	3	0	6
First	0	52	1	0	1	10	64
Lower Second	0	1	58	0	13	16	88
Pass	1	0	1	1	6	0	9
Third	0	1	9	0	82	12	104
Upper Second	0	8	11	0	5	111	135
Σ	4	62	80	1	110	149	406

Table 4.18 RF Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.946	0.261

The result of applying RF when including CAR only, to predict B Average Marks from Level A Average Marks, has a CA value of 0.946 and an RMSE of 0.261(as shown in Table 4.18).

Table 4.19 RF Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Predicted

Actual	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	8	1	5	0	0	4	18
First	0	97	0	0	0	14	111
Lower Second	0	3	53	0	4	10	70
Pass	0	0	1	1	1	1	4
Third	1	0	1	0	30	6	38
Upper Second	0	9	4	0	1	151	165
Σ	9	110	64	1	36	186	406

Table 4.20 RF Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.966	0.219

The result of applying RF when including CAR only, to predict C Average Marks from Level A and B Average Marks, has a CA value of 0.966 and an RMSE of 0.219 (as shown in Table 4.20).

Table 4.21 RF Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	7	3	2	0	1	5	18
First	1	84	3	0	1	22	111
Lower Second	1	3	44	0	3	19	70
Pass	0	0	1	1	0	2	4
Third	1	1	4	0	23	9	38
Upper Second	1	13	7	0	4	140	165
Σ	11	104	61	1	32	197	406

Table 4.22 RF Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.941	0.275

The result of applying RF when including CAR only, to predict C Average Marks from Level A Average Marks, has a CA value of 0.941 and an RMSE of 0.275 (as shown in table 4.22).

4.6.1.4 Applying RF Technique (Without SAC & CAR):

In this section, the RF technique was applied to the data, excluding SAC and CAR. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks

from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.23.

Table 4.23 RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	2	0	2	0	1	1	6
First	0	52	2	0	3	7	64
Lower Second	0	1	66	1	12	8	88
Pass	0	0	1	3	5	0	9
Third	0	2	4	0	92	6	104
Upper Second	0	7	6	0	10	112	135
Σ	2	62	81	4	123	134	406

Table 4.24 RF Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.894	0.469

The result of applying RF when excluding CAR and SAC, to predict B Average Marks from Level A Average Marks has a low CA value of 0.894 and an RMSE of 0.469 (as shown in Table 4.24).

Table 4.25 RF Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	11	0	3	0	0	4	18
First	0	106	1	0	0	4	111
Lower Second	1	1	55	0	1	12	70
Pass	1	0	1	0	0	2	4
Third	0	0	2	0	28	8	38
Upper Second	0	5	4	0	0	156	165
Σ	13	112	66	0	29	186	406

Table 4.26 RF Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Random Forest	0.936	0.284

The result of applying RF when excluding CAR and SAC, to predict C Average Marks from Level A and B Average Marks, has a low CA value of 0.936 and an RMSE of 0.284 (as shown in Table 4.26).

Table 4.27 RF Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	7	0	5	0	1	5	18
First	1	93	5	0	0	12	111
Lower Second	0	6	46	0	1	17	70
Pass	1	0	1	1	0	1	4
Third	1	4	7	0	18	8	38
Upper Second	1	14	6	0	3	141	165
Σ	11	117	70	1	23	184	406

Table 4.28 RF Evaluation Results of Predicted C Average Marks from Level A Average Marks

	Evaluation Results	
Method	**CA**	**RMSE**
Random Forest	0.897	0.346

The result of applying RF when excluding CAR and SAC, to predict C Average Marks from Level A Average Marks, has a low CA value of 0.897 and an RMSE of 0.346 (as shown in Table 4.28).

4.6.1.5 Summary of RF Technique Implementations:

The RF technique was applied to the data in four different scenarios. In terms of using SAC and CAR in predicting students' mark, each scenario had three

different options: firstly to predict B Average Marks from year A Average Marks, secondly to predict C Average Marks from year A and B Average Marks and thirdly to predict C Average Marks from year A Average Marks.

In general, it can be concluded that using SAC and CAR has a significant and positive effect on the CA, for all scenarios. However, further questions need to asked, for example:

- Why do we have different results for each level and what is the difference between using SAC only and CAR only?
- Does the RF technique have different results, compared with other proposed DM techniques?
- Is there a significant difference between using SAC and/or CAR and never using either of them?

These questions will be addressed in the following chapter.

4.6.2 Artificial Neural Networks (ANN):

In this section, applying Artificial Neural Networks (ANN) to the data is described. As highlighted in the literature, ANNs have been used for classification, regression, prediction, and other DM problems.

An ANN is one of the machine learning algorithms that can be used in different areas due to its flexibility and efficiency. "ANNs are a set of connected input/output units, usually called nodes. The output of all the nodes is considered an input to the next nodes. The output value of a node is, in

general, a non-linear function (referred to as the activation function) of its input value" (Wieland, et al. 2002). The multiplicative weighing factor between the input of node j and the output of node i is called the weight wji (shown in Figure 4.3).

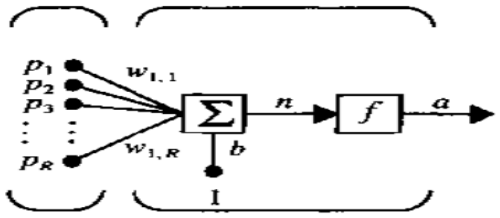


Figure 4.3: General Component of ANN Bishop (1995)

An ANN has an adaptive characteristic, which means that the values of each parameter or node are altered through an operation, generally called the Learning and Training stage. In the Training stage, ANN parameters are static and the system is arranged to solve the problems. During the Learning stage, the process of adjusting the weights to be able to predict the accurate class label of the input tuples in the first level of the ANN.

ANN has powerful tools for models, and it is also considered as a black box that has great capacity in predictive modelling. Research in this field of knowledge has led to the enhancement of many types of ANN, that are useful for solving various types of problems (Wieland et al. 2002).

ANNs are characterized by their capability of producing arbitrary, complex relationships between inputs and outputs. They are also able to analyze and organize data using intrinsic features, without any external guidance.

Additionally, ANNs of various kinds can be used for clustering and prototype creation.

On the other hand, ANNs do not work well when there are many hundreds or thousands of input features. They also do not yield acceptable performance for complex problems. (Yahia and El-mukashfi, 2010).

The ANN technique was applied to the data (as shown in Figure 4.4) using the Orange Canvas tool. The following four sections will represent the implementations of the ANN technique, in terms of using SAC and CAR in predicting students' marks. Each section has three different scenarios of data input, as mentioned in the previous section. Later, in Chapter Five, all results will be interpreted and evaluated in detail.

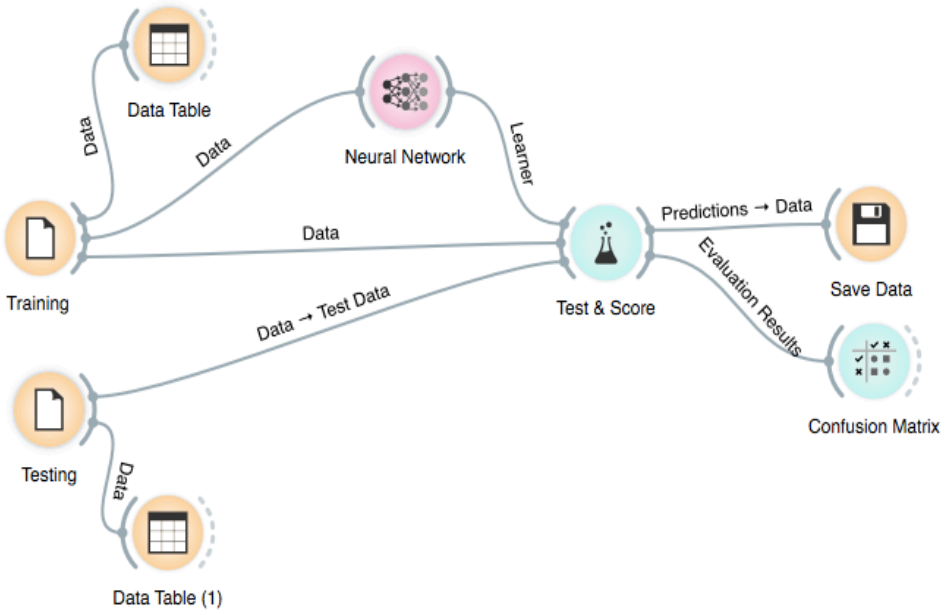


Figure 4.4 Applying ANN Technique on the Data

4.6.2.1 Applying Artificial Neural Network (ANN) Technique (With SAC & CAR):

In this section, the ANN technique was applied to the data, including SAC and Coursework CAR. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.29.

Table 4.29 ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	0	0	0	6	0	6
First	0	38	0	0	0	26	64
Lower Second	0	1	27	0	18	42	88
Pass	0	0	0	0	8	1	9
Third	0	0	16	0	61	27	104
Upper Second	0	20	14	0	15	86	135
Σ	0	59	57	0	108	182	406

Table 4.30 ANN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.978	0.173

The result of applying ANN when including CAR and SAC, to predict B Average Marks from Level A Average Marks, has a high CA of 0.978 and a low RMSE of 0.173 (as shown in Table 4.30).

Table 4.31 ANN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	2	0	6	0	6	4	18
	First	0	78	0	0	0	33	111
	Lower Second	0	2	20	0	3	45	70
	Pass	0	1	1	0	0	2	4
	Third	0	0	8	0	13	17	38
	Upper Second	0	16	11	0	1	137	165
	Σ	2	97	46	0	23	238	406

Table 4.32 ANN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.914	0.311

The result of applying ANN when including CAR and SAC, to predict C Average Marks from Level A and B Average Marks, has a CA value of 0.914 and an RMSE of 0.311 (as shown in Table 4.32).

Table 4.33 ANN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	2	0	10	0	1	5	18
First	0	77	1	0	0	33	111
Lower Second	0	5	21	0	0	44	70
Pass	0	0	1	0	0	3	4
Third	1	0	15	0	2	20	38
Upper Second	0	24	19	0	0	122	165
Σ	3	106	67	0	3	227	406

Table 4.34 ANN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.909	0.308

The result of applying ANN when including CAR and SAC, to predict C Average Marks from Level A Average Marks, has a CA of 0.909 and an RMSE of 0.308 (as shown in Table 4.34).

4.6.2.2 Applying Artificial Neural Network (ANN) Technique (With SAC only):

In this section, the ANN technique was applied to the data, including SAC only. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.35.

Table 4.35 ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	0	0	0	0	6	0	6
	First	0	50	0	0	0	14	64
	Lower Second	0	1	11	0	35	41	88
	Pass	0	0	0	0	9	0	9
	Third	0	0	6	0	75	23	104
	Upper Second	0	26	4	0	24	81	135
	Σ	0	77	21	0	149	159	406

Table 4.36 ANN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.899	0.352

The result of applying ANN when including SAC only, to predict B Average Marks from Level A Average Marks, has a CA value of 0.899 and an RMSE of 0.352 (as shown in Table 4.36).

Table 4.37 ANN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	2	0	3	0	8	5	18
	First	0	84	0	0	0	27	111
	Lower Second	0	4	5	0	8	53	70
	Pass	0	1	1	0	0	2	4
	Third	1	0	6	0	13	18	38
	Upper Second	0	22	4	0	0	139	165
	Σ	3	111	19	0	29	244	406

Table 4.38 ANN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.899	0.347

The result of applying ANN when including SAC only, to predict C Average Marks from Level A and B Average Marks, has a CA of 0.899 and an RMSE of 0.347 (as shown in Table 4.38).

Table 4.39 ANN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	5	0	1	0	4	8	18
First	0	75	0	0	0	36	111
Lower Second	1	5	9	0	3	52	70
Pass	0	0	0	0	0	4	4
Third	2	0	2	0	6	28	38
Upper Second	0	21	2	0	0	142	165
Σ	8	101	14	0	13	270	406

Table 4.40 ANN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.904	0.331

The result of applying ANN when including SAC only, to predict C Average Marks from Level A and B Average Marks, has a CA of 0.904 and an RMSE of 0.331 (as shown in Table 4.40).

4.6.2.3 Applying Artificial Neural Network (ANN) Technique (With CAR only):

In this section, the ANN technique was applied to the data, including CAR only. Three different scenarios were applied, firstly to predict B Average

Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.41.

Table 4.41 ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	0	0	1	0	5	0	6
	First	0	46	0	0	0	18	64
	Lower Second	0	1	19	0	31	37	88
	Pass	0	0	1	0	7	1	9
	Third	0	1	17	0	64	22	104
	Upper Second	0	29	12	0	19	75	135
	Σ	0	77	50	0	126	153	406

Table 4.42 ANN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.879	0.387

The result of applying ANN when including CAR only, to predict B Average Marks from Level A Average Marks, has a CA of 0.879 and an RMSE of 0.387 (as shown in Table 4.42).

Table 4.43 ANN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	2	7	0	1	8	18
First	0	76	0	0	0	35	111
Lower Second	0	5	16	0	1	48	70
Pass	0	0	1	0	0	3	4
Third	0	0	7	0	3	28	38
Upper Second	0	24	4	0	0	137	165
Σ	0	107	35	0	5	259	406

Table 4.44 ANN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.909	0.314

The result of applying ANN when including CAR only, to predict C Average Marks from Level A and B Average Marks, has a CA value of 0.909 and an RMSE of 0.314 (as shown in Table 4.44).

Table 4.45 ANN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	1	2	0	0	15	18
First	0	70	0	0	0	41	111
Lower Second	0	7	0	0	0	63	70
Pass	0	0	0	0	0	4	4
Third	0	0	1	0	0	37	38
Upper Second	0	30	1	0	0	134	165
Σ	0	108	4	0	0	294	406

Table 4.46 ANN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.906	0.306

The result of applying ANN when including CAR only, to predict C Average Marks from Level A Average Marks, has a CA of 0.906 and an RMSE of 0.306 (as shown in Table 4.46).

4.6.2.4 Applying Artificial Neural Network (ANN) Technique (Without SAC & CAR):

In this section, the ANN technique was applied to the data, excluding SAC and CAR. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks

from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.47.

Table 4.47 ANN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	0	1	0	5	0	6
First	0	34	0	0	0	30	64
Lower Second	0	1	21	0	25	41	88
Pass	0	0	1	0	7	1	9
Third	0	0	19	0	60	25	104
Upper Second	0	16	8	0	17	94	135
Σ	0	51	50	0	114	191	406

Table 4.48 ANN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.759	0.51

The result of applying ANN when excluding CAR and SAC, to predict B Average Marks from Level A Average Marks, has a very low CA with a value of 0.759 and an RMSE of 0.51 (as shown in Table 4.48).

Table 4.49 ANN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	2	11	0	0	5	18
First	0	78	1	0	0	32	111
Lower Second	0	4	29	0	0	37	70
Pass	0	0	1	0	0	3	4
Third	0	0	19	0	0	19	38
Upper Second	0	25	12	0	0	128	165
Σ	0	109	73	0	0	224	406

Table 4.50 ANN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.791	0.461

The result of applying ANN when excluding CAR and SAC, to predict C Average Marks from Level A and B Average Marks, has a very low CA value of 0.791 and an RMSE of 0.461 (as shown in Table 4.50).

Table 4.51 ANN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	1	0	0	0	17	18
First	0	71	0	0	0	40	111
Lower Second	0	7	0	0	0	63	70
Pass	0	0	0	0	0	4	4
Third	0	0	0	0	0	38	38
Upper Second	0	25	0	0	0	140	165
Σ	0	104	0	0	0	302	406

Table 4.52 ANN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
Artificial Neural Network	0.820	0.44

The result of applying ANN when excluding CAR and SAC, to predict C Average Marks from Level A Average Marks, has a very low CA of 0.820 and an RMSE of 0.44 (as shown in Table 4.52).

4.6.2.5 Summary of ANN Implementations:

ANN was applied to the data in four different scenarios, in terms of using SAC and CAR in predicting students' marks. Each scenario has three different options: firstly to predict B Average Marks from year A Average

Marks, secondly to predict C Average Marks from year A and B Average Marks and thirdly to predict C Average Marks from year A Average Marks.

In general, it can be concluded that using SAC and CAR have a significant and positive effect on the CA, for all scenarios. However, further questions need to be asked, for example:

- Why do we have different results for each level?
- What is the difference between using SAC only and CAR only?
- Does the ANN technique have different results, compared with other proposed DM techniques?
- Is there a significant difference between using SAC and/or CAR and never using either of them?

These questions will be addressed in the following.

4.6.3 k-Nearest Neighbors (kNN)

In this section, applying k-Nearest Neighbors (kNN) to the data is described. As highlighted in the literature, kNN have been used for classification, decision rules, prediction, and other DM problems.

The kNN classifier is a conventional non-parametric classifier that gives high performance for optimal values of k. In the kNN rule, a test sample is choosing the class that represented frequently among the k nearest training samples. When two classes or more exist, the test sample will choose the class with closest distance to it.

It can be shown that the k-nearest neighbor rule becomes the Bayes optimal decision rule, as k goes to infinity (Duda and Hart, 1973). It is only in the limit, as the number of training samples goes to infinity, that the nearly optimal behavior of the k-nearest neighbor rule is assured.

The kNN technique was applied to the data (as shown in Figure 4.5) using the Orange Canvas tool. The following four sections will represent the implementations of the kNN technique, in terms of using SAC and CAR in predicting students' marks. Each section has three different scenarios of data input, as mentioned in the previous section. Later in Chapter Five, all results will be interpreted and evaluated in detail.

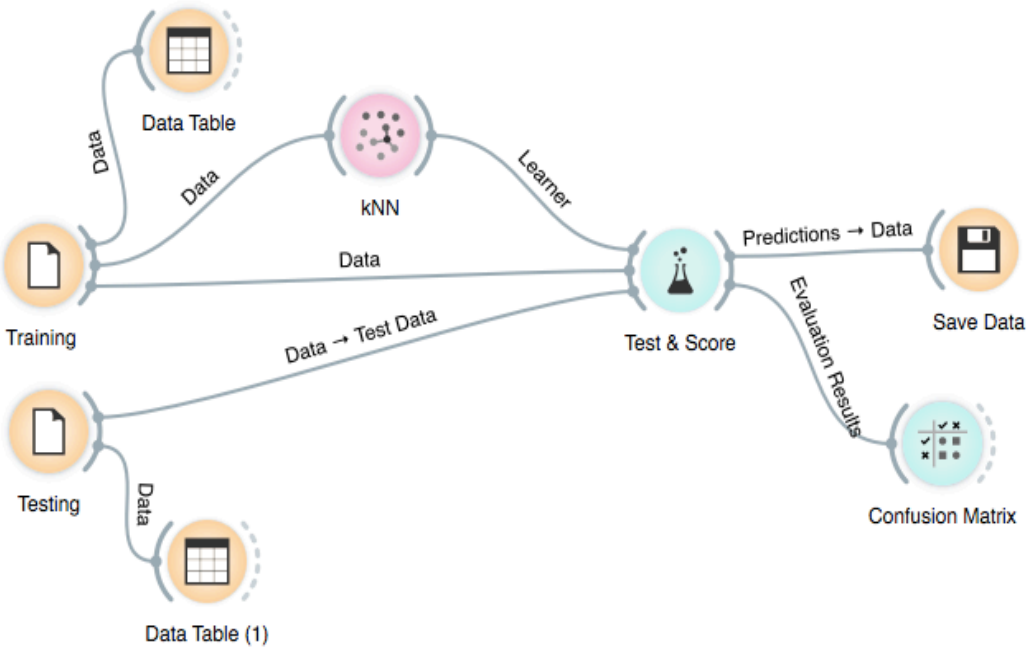


Figure 4.5. Applying k-NN Technique on the data

4.6.3.1 Applying k-Nearest Neighbors (kNN) Technique (With SAC & CAR):

In this section, the kNN technique was applied to the data, including SAC and CAR. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.53.

Table 4.53 kNN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	0	0	1	0	5	0	6
	First	0	45	2	0	3	14	64
	Lower Second	0	3	56	0	15	14	88
	Pass	0	0	1	0	8	0	9
	Third	0	1	23	0	71	9	104
	Upper Second	0	24	16	0	16	79	135
	Σ	0	73	99	0	118	116	406

Table 4.54 kNN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.978	0.173

The result of applying the kNN technique when including CAR and SAC, to predict B Average Marks from Level A Average Marks, has a CA value of 0.978 and an RMSE of 0.173 (as shown in Table 4.54).

Table 4.55 kNN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	7	4	3	0	1	3	18
	First	1	90	3	0	0	17	111
	Lower Second	5	7	34	0	7	17	70
	Pass	1	3	0	0	0	0	4
	Third	3	1	12	0	12	10	38
	Upper Second	1	26	15	0	4	119	165
	Σ	18	131	67	0	24	166	406

Table 4.56 kNN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.906	0.359

The result of applying the kNN technique when including CAR and SAC, to predict C Average Marks from Level A and B Average Marks, has a CA of 0.906 and an RMSE of 0.359 (as shown in table 4.56).

Table 4.57 kNN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

		Predicted						
		Fail	**First**	**Lower Second**	**Pass**	**Third**	**Upper Second**	**Σ**
Actual	**Fail**	6	4	2	0	2	4	**18**
	First	1	95	3	0	1	11	**111**
	Lower Second	3	10	31	0	1	25	**70**
	Pass	0	2	0	0	0	2	**4**
	Third	1	3	9	0	12	13	**38**
	Upper Second	2	32	19	0	5	107	**165**
	Σ	**13**	**146**	**64**	**0**	**21**	**162**	**406**

Table 4.58 kNN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.914	0.296

The result of applying the kNN technique when including CAR and SAC, to predict C Average Marks from Level A Average Marks, has a CA value of 0.914 and an RMSE of 0.296 (as shown in Table 4.58).

4.6.3.2 Applying k-Nearest Neighbors (kNN) Technique (With SAC only):

In this section, the kNN technique was applied to the data, including SAC only. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.59.

Table 4.59 kNN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	0	0	1	0	4	1	6
	First	0	47	1	0	1	15	64
	Lower Second	0	3	57	0	17	11	88
	Pass	0	0	1	0	8	0	9
	Third	1	1	19	0	73	10	104
	Upper Second	0	27	15	0	15	78	135
	Σ	1	78	94	0	118	115	406

Table 4.60 kNN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.882	0.375

The result of applying the kNN technique when including SAC only, to predict B Average Marks from Level A Average Marks, has a CA value of 0.882 and an RMSE of 0.375 (as shown in Table 4.60).

Table 4.61 kNN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	7	4	3	0	1	3	18
	First	1	90	3	0	0	17	111
	Lower Second	5	7	35	0	7	16	70
	Pass	1	3	0	0	0	0	4
	Third	3	1	12	0	12	10	38
	Upper Second	1	25	15	0	4	120	165
	Σ	18	130	68	0	24	166	406

Table 4.62 kNN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.906	0.317

The result of applying the kNN technique when including SAC only, to predict C Average Marks from Level A and B Average Marks has, a CA value of 0.906 and an RMSE of 0.317 (as shown in Table 4.62).

Table 4.63 kNN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	5	3	2	0	3	5	18
	First	1	99	2	0	1	8	111
	Lower Second	3	12	27	0	0	28	70
	Pass	0	2	1	0	0	1	4
	Third	1	4	8	0	10	15	38
	Upper Second	2	35	18	0	6	104	165
	Σ	12	155	58	0	20	161	406

Table 4.64 kNN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.906	0.316

The result of applying the kNN technique when including SAC only, to predict C Average Marks from Level A Average Marks, has a CA value of 0.906 and an RMSE of 0.316 (as shown in Table 4.64).

4.6.3.3 Applying k-Nearest Neighbors (kNN) Technique (With CAR only):

In this section, the kNN technique was applied to the data, including CAR only. Three different scenarios were applied, firstly to predict B Average

Marks from level A Average Marks, secondly to predict C Average Marks from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.65.

Table 4.65 kNN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	0	0	3	0	3	0	6
First	0	44	2	0	3	15	64
Lower Second	0	3	56	0	16	13	88
Pass	0	0	2	0	7	0	9
Third	0	1	27	0	67	9	104
Upper Second	1	21	20	0	12	81	135
Σ	1	69	110	0	108	118	406

Table 4.66 kNN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.889	0.353

The result of applying the kNN technique when including CAR only, to predict B Average Marks from Level A Average Marks, has a CA value of 0.889 and an RMSE of 0.353 (as shown in Table 4.66).

Table 4.67 kNN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	7	4	3	0	1	3	18
First	1	90	3	0	0	17	111
Lower Second	5	7	34	0	7	17	70
Pass	1	3	0	0	0	0	4
Third	3	1	12	0	12	10	38
Upper Second	1	26	15	0	4	119	165
Σ	18	131	67	0	24	166	406

Table 4.68 kNN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest neighbors	0.906	0.319

The result of applying the kNN technique when including CAR only, to predict C Average Marks from Level A and B Average Marks, has a CA value of 0.906 and an RMSE of 0.319 (as shown in Table 4.68).

Table 4.69 kNN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	3	2	3	0	1	9	18
First	1	93	2	0	1	14	111
Lower Second	3	12	31	0	3	21	70
Pass	0	1	0	0	0	3	4
Third	1	5	10	0	10	12	38
Upper Second	1	29	22	0	3	110	165
Σ	9	142	68	0	18	169	406

Table 4.70 kNN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.911	0.309

The result of applying the kNN technique when including CAR only, to predict C Average Marks from Level A Average Marks, has a CA value of 0.911 and an RMSE of 0.309 (as shown in Table 4.70).

4.6.3.4 Applying k-Nearest Neighbors (kNN) Technique (Without SAC & CAR):

In this section, the kNN technique was applied to the data, excluding CAR and SAC. Three different scenarios were applied, firstly to predict B Average Marks from level A Average Marks, secondly to predict C Average Marks

from level A and B Average Marks and thirdly to predict C Average Marks from level A Average Marks. The RF Confusion Matrix of Predicted B Average Marks from Level A Average Marks is given in Table 4.71.

Table 4.71 kNN Confusion Matrix of Predicted B Average Marks from Level A Average Marks

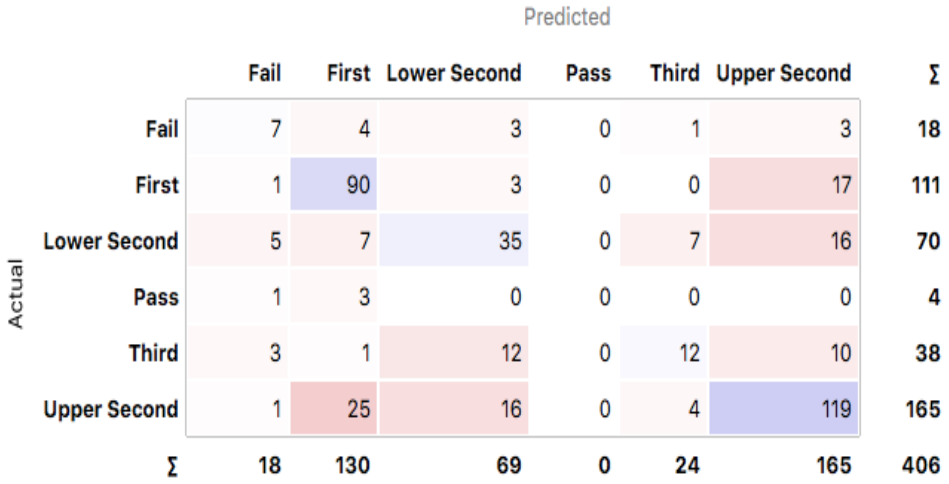
Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	1	0	3	0	1	1	6
First	0	46	3	0	1	14	64
Lower Second	1	3	56	0	16	12	88
Pass	0	0	2	0	7	0	9
Third	1	3	23	0	68	9	104
Upper Second	1	26	21	0	14	73	135
Σ	4	78	108	0	107	109	406

Table 4.72 kNN Evaluation Results of Predicted B Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest Neighbors	0.815	0.441

The result of applying the kNN technique when excluding CAR and SAC, to predict B Average Marks from Level A Average Marks, has a low CA value of 0.815 and an RMSE of 0.441 (as shown in Table 4.72).

Table 4.73 kNN Confusion Matrix of Predicted C Average Marks from Level A and B Average Marks



Actual \ Predicted	Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Fail	7	4	3	0	1	3	18
First	1	90	3	0	0	17	111
Lower Second	5	7	35	0	7	16	70
Pass	1	3	0	0	0	0	4
Third	3	1	12	0	12	10	38
Upper Second	1	25	16	0	4	119	165
Σ	18	130	69	0	24	165	406

Table 4.74 kNN Evaluation Results of Predicted C Average Marks from Level A and B Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest neighbors	0.830	0.426

The result of applying the kNN technique when excluding CAR and SAC, to predict C Average Marks from Level A and B Average Marks, has a low CA value of 0.830 and an RMSE of 0.426 (as shown in Table 4.74).

Table 4.75 kNN Confusion Matrix of Predicted C Average Marks from Level A Average Marks

		Predicted						
		Fail	First	Lower Second	Pass	Third	Upper Second	Σ
Actual	Fail	4	2	3	0	0	9	18
	First	1	97	2	0	1	10	111
	Lower Second	3	15	30	0	2	20	70
	Pass	1	1	1	0	0	1	4
	Third	0	5	10	0	11	12	38
	Upper Second	1	36	24	0	3	101	165
	Σ	10	156	70	0	17	153	406

Table 4.76 kNN Evaluation Results of Predicted C Average Marks from Level A Average Marks

Evaluation Results		
Method	**CA**	**RMSE**
k-Nearest neighbors	0.820	0.438

The result of applying the kNN technique when excluding CAR and SAC, to predict C Average Marks from Level A Average Marks, has a low CA value of 0.820 and an RMSE of 0.438 (as shown in table 4.76).

4.6.3.5 Summary of k-NN Implementations:

The k-NN technique was applied to the data in four different scenarios, in terms of using SAC and CAR in predicting students' mark. Each scenario has

three different options; firstly to predict B Average Marks from year A Average Marks, secondly to predict C Average Marks from year A and B Average Marks and thirdly, to predict C Average Marks from year A Average Marks.

It can be concluded that using SAC and CAR has a significant and positive effect on the CA for all scenarios. However, the following questions need to be addressed:

- Why do we have different results for each level?
- What is the difference between using SAC only and CAR only?
- Does the k-Nearest Neighbors (kNN) technique have different results compared with other DM techniques?
- Is there a significant difference between using SAC and/or CAR and never using either of them?

These questions will be answered in detail in the next chapter.

4.7 Conclusions:

Data using DM techniques had been pre-processed through different stages. The stages included adding new attributes that increased the accuracy of prediction, as described in this chapter, and other various pre-processing stages to guarantee clean and manageable data.

After splitting the modified data into 70% for training and 30% for testing, the three DM techniques were trained on the data that contained records of students who completed their entire programme (i.e. including their Average A, B, C, SAC and CAR). After that, experiments of used DM techniques investigated the prediction of final students' marks of the second year, based on first year's marks, prediction of final students' marks of the third year, based on the first and second years' marks, and prediction of final students' marks of the third year, based on the first years' marks only.

Additionally, in terms of SAC and CAR in predicting students' marks, further investigation was carried out, such as, predicting students' average marks while excluding SAC, and in another experiment including it, predicting students' average marks while excluding CAR, and in another experiment including it, and predicting students' marks while including both CAR and SAC.

Finally, average SAC and CAR were removed from the dataset, which helped in determining the difference between using the newly formulated attributes (SAC and CAR) and not using them, in the prediction task. In chapter 4, the

fifth objective has been achieved successfully by showing the importance of applying the DM classification techniques on HE students' data to predict and classify students' academic performance, while taking into consideration the accuracy of input data.

In Chapter 5, results from the experiments in this chapter are compared with related studies that have the same objectives, factors, and DM techniques. The aim of this comparison is to ensure that the new preparation process will positively affect the final results. Finally, a detailed evaluation will be provided for all experiment results.

CHAPTER FIVE: RESULTS AND EVALUATION

5.1. Introduction:

This chapter describes the steps followed to achieve various predictive models of students' academic performance. More specifically, results of using three main DM techniques (Random Forest, k-Nearest Neighbours, and Artificial Neural Networks) will be interpreted and evaluated, including all experimental scenarios, in order to evaluate the results of final students' marks based on previous years' marks (second year based on first year's marks, third year based on the first and second years' marks, and results of final students' average marks based only on second and third years' marks).

In terms of SAC and CAR, in predicting students' marks, further evaluation will be carried out. Some scenarios are described below:

- Evaluating and comparing the results of students' average marks, when excluding SAC and including it.
- Evaluating and comparing the results of students' average marks, when excluding CAR and including it.
- Evaluating and comparing the results of students' average marks, when including both CAR and SAC.
- Evaluating and comparing the results of students' average marks, when excluding both CAR and SAC.

Additionally, this chapter will discuss the importance of using SAC and CAR to increase the prediction accuracy. The results will then be compared with related studies and research.

5.2. Data Mining (DM) Techniques - Results:

This section will present the results of applying DM techniques. However, each DM technique has four scenarios, in terms of SAC and CAR in predicting students' average marks. Each scenario has three different options, as mentioned in Chapter Four.

5.2.1. Random Forest (RF) Results:

The RF technique was applied to the data in four scenarios, in terms of SAC and CAR in predicting students' average marks. Each scenario has three different options: firstly to predict B Average Marks from year A Average Marks, secondly to predict C Average Marks from year A and B Average Marks and thirdly to predict C Average Marks from year A Average Marks. The results are given in Table 5.1 below:

Table 5.1. Random Forest (RF) Results

Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE
Year A to predict B	SAC & CAR	0.988	0.141
	SAC only	0.975	0.207
	CAR only	0.946	0.261
	Without SAC & CAR	0.894	0.469
Year A and B to predict C	SAC & CAR	0.973	0.184
	SAC only	0.961	0.205
	CAR only	0.966	0.219
	Without SAC & CAR	0.936	0.285
Year A to predict C	SAC & CAR	0.963	0.228
	SAC only	0.958	0.221
	CAR only	0.941	0.276
	Without SAC & CAR	0.897	0.346

As outlined in the previous chapter, the following questions have been formulated from the results that were generated using the RF technique:

1) What is the difference between using SAC only and CAR only?
2) Is there any significant difference between including SAC and CAR, and excluding them?
3) Why do we have different results for each level and which level would provide the best prediction results?
4) Does the RF technique have different results when compared with other DM techniques?

In terms of SAC and CAR, the following scenarios should answer the first two questions in detail:

- First Scenario: Results of predicting final students' average marks of the second year, based on the first year's marks, showing that when SAC and CAR are included, the CA value of 0.988 and RMSE value of 0.141, gave the highest accuracy and lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results still show high values for CA and low values for RMSE. In this case, using SAC only gave better results, compared with using CAR only. In other words, including SAC only has affected the results by increasing the CA and decreasing the

RMSE, compared with using CAR only. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.894), and the highest RMSE (with a value of 0.469), compared with including both CAR and SAC or including one of them. In other words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

- Second Scenario: Results of predicting final students' marks of the third year, based on the first and second years' marks, show that when SAC and CAR are included, the CA has a value of 0.973 and RMSE has a value of 0.184, giving the highest accuracy and lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave almost the same results, compared with using CAR only. In other words, including SAC only or CAR only, has affected the results by increasing the CA and decreasing the RMSE. However, including SAC and CAR together, has the maximum effect on the results, compared with using SAC or CAR separately.

In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.936), and the highest RMSE (with a value of 0.285), compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the classification accuracy (CA) and decreasing the RMSE, compared with excluding SAC and CAR.

- Third Scenario: Results of predicting final students' marks of the third year, based on the first years' marks only, show that when SAC and CAR are included, the results of CA (with a value of 0.963) and RMSE (with a value of 0.228), gave the highest accuracy and lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset, has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave better results, compared with using CAR only. In other words, including SAC only, has affected the results by increasing the CA and decreasing the RMSE, compared with using CAR only. However, including SAC and CAR together, has the maximum effect on the results, compared with using SAC or CAR separately.

  In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.897), and the highest RMSE (with a value of 0.346), compared with including both CAR and SAC or including one of them. In other

words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

In terms of comparing all scenarios together using RF, the third question presented above, can now be answered. It has been found that the results of the first scenario; predicting final students' marks of the second year, based on the first year's marks, give the best results. In other words, predicting final students' marks of the second year, based on the first year's marks, gives the best indication, in terms of students' final mark prediction, compared with other scenarios. However, the reason for this is linked to the amount of data in the experiment and it was mentioned before that many students may study one or two years only, or they may continue their entire program and study all three years. This was reflected in the final modified data, where the number of records for students of level A (788 records) was greater than both students of level B (606 records) and students of level C (406 records). In addition, the RF technique works better with bigger sizes of data (Breiman, 2001), in other words, a bigger dataset, gives better classification accuracy.

5.2.2. Artificial Neural Networks (ANN) Results:

The ANN technique was applied to the data in four scenarios, in terms of using SAC and CAR in predicting students' marks. Each scenario has three different options: firstly to predict B Average Marks from year A Average Marks, secondly to predict C Average Marks from year A and B Average

Marks and thirdly to predict C Average Marks from year A Average Marks. Experiment results are given in Table 5.2.

Table 5.2. Artificial Neural Networks (ANN) Results

Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE
Year A to predict B	SAC & CAR	0.978	0.173
	SAC only	0.899	0.352
	CAR only	0.879	0.387
	Without SAC & CAR	0.759	0.511
Year A and B to predict C	SAC & CAR	0.914	0.311
	SAC only	0.899	0.348
	CAR only	0.909	0.315
	Without SAC & CAR	0.791	0.462
Year A to predict C	SAC & CAR	0.909	0.308
	SAC only	0.904	0.332
	CAR only	0.906	0.307
	Without SAC & CAR	0.820	0.440

According to the results of using the ANN technique (as shown in Table 5.2), some important questions must be addresses:

1) What is the difference between using SAC only and CAR only?
2) Is there any significant difference between including SAC and CAR and excluding them?
3) Why do we have different results for each level and which level could make better prediction results than other?
4) Does ANN have different results, compared with other DM techniques?

In terms of SAC and CAR, the following scenarios should answer the first two questions in detail:

- First Scenario: Results of predicting final students' marks of the second year, based on first year's marks, show that when SAC and CAR are included, CA has a value of 0.978 and an RMSE value of 0.173, giving the highest accuracy and lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave better results, compared with using CAR only. In other words, including SAC only has affected the results by increasing the CA and decreasing the RMSE, compared with using CAR only. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

In terms of excluding SAC and CAR, the results show the lowest CA (with value 0.759), and the highest RMSE (with value 0.511), compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

- Second Scenario: Results of predicting final students' marks of the third year, based on the first and second years' marks, show that when SAC and CAR are included, CA has a value of 0.914 and an RMSE value of 0.311, giving the highest CA and the lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using CAR only gave better results, compared with using SAC only. In other words, including CAR only has affected the results by increasing the CA and decreasing the RMSE, compared with using SAC only. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

  In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.791) and the highest RMSE (with a value of 0.462), compared with including both CAR and SAC, or including one of them. In other

words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

- Third Scenario: Results of predicting final students' marks of the third year, based on the first year's marks only, show that when SAC and CAR are included, the results of CA (with a value of 0.909) and RMSE (with a value of 0.308) gave the highest CA and the lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset, has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave almost the same results, compared with using CAR only. In other words, including SAC only or CAR only, has affected the results by increasing the CA and decreasing the RMSE. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

  In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.820), and the highest RMSE (with a value of 0.440), compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

In terms of comparing all scenarios together using ANN, the third question (regarding the reasons for different results for each level and which level could lead to better prediction results), should be answered here. It has been found that the results of the first scenario (predicting final students' marks of the second year, based on the first year's marks), give the best results. In other words, predicting final students' marks of the second year, based on first year's marks, gives the best indication, in terms of students' final marks prediction, compared with other scenarios.

The ANN technique is a supervised learning machine that also considers the size of the data. However, it was mentioned before that the number of records for students of level A (788 records) is greater than both students of level B (606 records) and students of level C (406 records). Similarly to the RF technique, ANN works better with bigger amounts of data, which is why predicting final students' marks of the second year, based on the first year's marks, gives better CA results and is the best indication, in terms of students' final marks prediction, compared with other scenarios.

5.2.3. k-Nearest Neighbours (kNN) Results:

The kNN technique was applied to the data in four scenarios in terms of using SAC and CAR in predicting students' marks. Each scenario has three different options: firstly to predict B Average Marks from year A Average Marks, secondly to predict C Average Marks from year A and B Average Marks and thirdly to predict C Average Marks from year A Average Marks. The results are given in Table 5.3.

Table 5.3. k-Nearest Neighbours (kNN) Results

Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE
Year A to predict B	SAC & CAR	0.978	0.173
	SAC only	0.882	0.375
	CAR only	0.889	0.354
	Without SAC & CAR	0.815	0.442
Year A and B to predict C	SAC & CAR	0.906	0.359
	SAC only	0.906	0.318
	CAR only	0.906	0.319
	Without SAC & CAR	0.830	0.427
Year A to predict C	SAC & CAR	0.914	0.297
	SAC only	0.906	0.316
	CAR only	0.911	0.310
	Without SAC & CAR	0.820	0.438

According to the results of using the kNN technique (as shown in Table 5.3), questions outlined in the previous chapter should be answered here:

1) What is the difference between using SAC only and CAR only?
2) Is there any significant difference between including SAC and CAR, and excluding them?
3) Why do we have different results for each level and which level could make better prediction results?
4) Does kNN have different results, compared with other DM techniques?

In terms of SAC and CAR, the following scenarios should answer the first two questions in detail:

- First Scenario: Results of predicting final students' marks of the second year, based on the first year's marks, show that when SAC and CAR are included, the results of CA, with a value of 0.978 and an RMSE value of 0.173, gives the highest CA and the lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset, has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave almost the same results, compared with using CAR only. In other words, including SAC only or CAR only, has affected the results by increasing the CA and decreasing the RMSE. However, including SAC and CAR together has the

maximum effect on the results, compared with using SAC or CAR separately.

In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.815), and the highest RMSE (with a value of 0.442), compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

- Second Scenario: Results of predicting final students' marks of the third year, based on the first and second year's marks, show that when SAC and CAR are included, the results of CA (with a value of 0.906) and RMSE (with a value of 0.359), gave the highest CA and the lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset, has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using SAC only gave the same results, compared with using CAR only. In other words, including SAC only or CAR only, has affected the results by increasing the CA and decreasing the RMSE. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.830) and the highest RMSE (with a value of 0.427), compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

- Third Scenario: Results of predicting final students' marks of the third year, based on the first year's marks only, show that when SAC and CAR are included, the results of CA (with a value of 0.914) and RMSE (with a value of 0.297), gave the highest CA and the lowest RMSE, compared with other terms of using SAC and CAR. In other words, including SAC and CAR in the dataset, has affected the results by increasing the CA and decreasing the RMSE.

  In terms of using SAC only or CAR only, the results show high values for CA and low values for RMSE. In this case, using CAR only gave better results, compared with using SAC only. In other words, including CAR only has affected the results by increasing the CA and decreasing the RMSE, compared with using SAC only. However, including SAC and CAR together has the maximum effect on the results, compared with using SAC or CAR separately.

  In terms of excluding SAC and CAR, the results show the lowest CA (with a value of 0.820) and the highest RMSE (with a value of 0.438), compared with including both CAR and SAC, or including one of them. In other

words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR.

In terms of comparing all scenarios together, using kNN, it has been found that the results of the first scenario give the best indication, in terms of students' final marks prediction, compared with other scenarios. The kNN technique is a supervised learning machine that also considers the size of data. Similarly to other supervised learning machines, kNN works better with bigger amounts of data, which is why predicting final students' marks of the second year, based on the first year's marks, gives better CA results and the best indication, in terms of students' final marks prediction, compared with other scenarios. In addition, the second scenario gives a better CA, compared with the third scenario, for the same reason.

5.3. Comprehensive Results Evaluation of DM Techniques:

From the results discussed in the previous section, the first scenario of the three DM techniques has been chosen for evaluation. The first scenario has the highest CA values and the lowest RMSE values for all DM techniques used in this thesis. All DM technique results are given in Table 5.4.

Table 5.4. All Data Mining Techniques: Evaluation Results

	Random Forest (RF) Technique		
Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE
Year A to predict B	SAC & CAR	0.988	0.141
	SAC only	0.975	0.207
	CAR only	0.946	0.261
	Without SAC & CAR	0.894	0.469
	Artificial Neural Networks (ANN) Technique		
Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE
Year A to predict B	SAC & CAR	0.978	0.173
	SAC only	0.899	0.352
	CAR only	0.879	0.387
	Without SAC & CAR	0.759	0.511
	k-Nearest Neighbours (kNN) Technique		
Years Prediction Scenarios	Terms of Using SAC and CAR	Evaluation Results	
		CA	RMSE

Year A to predict B	SAC & CAR	0.978	0.173
	SAC only	0.882	0.375
	CAR only	0.889	0.354
	Without SAC & CAR	0.815	0.442

According to the results shown in Table 5.4, it can be concluded that the RF technique gave the best results; the highest CA with a value of 0.988, and the lowest RMSE with a value of 0.14,1 compared with other DM techniques, in terms of using SAC and CAR. RF and ANN techniques followed kNN.

In terms of using SAC only or CAR only, the results still show high values for CA and low values for RMSE, for all DM techniques. In this case, using SAC only, gave better results, compared with using CAR only. In other words, including SAC only has affected the results by increasing the CA and decreasing the RMSE, compared with using CAR only. However, including SAC and CAR together have the maximum effect on the results, compared with using SAC or CAR separately, and this is true for all DM techniques.

In terms of excluding SAC and CAR, the lowest CA was a value of 0.894, and the highest RMSE was a value of 0.469 for RF. For ANN the lowest CA was a value of 0.759, and highest RMSE was a value of 0.511 and for kNN the lowest CA was a value of 0.815 and the highest RMSE was a value of 0.442, compared with including both CAR and SAC, or including one of them. In other words, including SAC and CAR, or one of them, in the dataset, has affected the results by increasing the CA and decreasing the RMSE, compared with excluding SAC and CAR, for all DM techniques used in this thesis.

Now, it is important to understand why the RF technique gave better results than the ANN and kNN techniques, and why the ANN technique was a better classifier than kNN. In fact, in terms of applying different DM techniques, there are some factors which may affect the CA, such as data size, attributes' features, data type and status of data (if they are prepared or not).

The RF technique gave the highest CA when compared with other DM techniques. In terms of the performance of the used DM techniques, the RF technique is faster than the ANN and kNN techniques and RF is also easier to train because it has less hyper-parameters to tune, knowing that the hyper-parameter is a parameter whose value is set before the learning process begins. On other hand, the values of other parameters are coming from training. In contrast, for ANN there are a huge number of parameters to choose from, such as number of layers, number of neurons in each layer, activation function and learning rate. However, RF is less prone to over-fit because it is designed from scratch to resist over-fitting.

In terms of data size, the RF technique can handle a small data size and give high accuracy but most ANN technique require a large amount of data to make generalizations, and the number of examples should be much larger than the number of features, whereas RF can provide better accuracy with only a small number of examples, even with too many features. Moreover, ANN is in many cases a black-box solution, which means that it is often very hard to understand what happens during classification or what features matter the most for the different classes. On other hand, using RF makes it much easier

to see which features are influencing the classification decision. RF has fewer parameters to optimize, it works with the number of estimators and the depth of each tree, whilst ANN has a learning rate, number of layers and width of layers, to name a few.

RF can be used for both classification and regression tasks. Over-fitting is one critical problem that may negatively impact the results, but with the RF technique, if there are enough trees in the forest, the classifier will not over-fit the model. Moreover, RF can handle missing values, and can be modeled for a categorical value that cannot be offered by the kNN technique.

It can be concluded that the RF technique would triumph over other DM techniques and have better CA in terms of the used data and their types, exist attributes' features and status of the data preparation in this thesis. The next section will discuss the importance of using SAC and CAR in increasing the prediction accuracy.

5.4. Importance of SAC and CAR in Increasing Prediction Accuracy:

To determine the impact of SAC and CAR on prediction accuracy, the same DM techniques were applied to data, without considering SAC and CAR as attributes. In other words, in the previous experiments, RF, kNN, and ANN were applied to predict their year *B* average marks from their year *A* average marks, to predict their year *C* average marks from their year *A and B* average marks, and to predict their year *C* average marks from their year *A* average marks.

Different studies have discussed various methods through which students' academic performance is predicted, based on different attributes and features. In this thesis, two main objectives were achieved: firstly, predicting students' academic performance, measured by students' total average marks and their average SAC and average CAR. The three DM techniques, that were chosen based on comparing a set of the most common DM techniques used for predicting students' performance, were applied to prepared data. The results have shown that RF, kNN, and ANN were able to predict students' total average marks with variable values of CA and RMSE: RF being the most accurate and kNN the least.

The second main objective that was achieved in this thesis was comparing results achieved from the first objective, with results from applying the same three DM techniques on the same data, but without considering average SAC and CAR. In other words, would the attributes Average SAC and Average CAR increase the accuracy of the prediction?

By applying the DM techniques, prediction accuracy fell dramatically and errors increased with RMSE, varying from the RF's 0.469, kNN's 0.442, and finally ANN's worst RMSE of 0.511 (as shown in Table 5.4).

The results also show that the formulated Student Attendance (represented by SAC) and the Coursework Assessment Ratio (represented by CAR) have an impact on both students' final results, as shown in the literature, and on the accuracy of predicting these marks.

5.5. Comparing Results with Other Related Work:

In this section, related work will be investigated and compared with the results of this thesis. However, previous studies have been concerned with students' attendance for estimating their performance, without investigating the various factors that may affect the credibility of students' attendance itself. In other words, it appears to be that no studies in the literature have considered looking at another form of the students' attendance, rather than relying on its classical form (i.e. number of attendance occurrences) (Brijesh et al., 2011).

By reviewing the literature that considers students' attendance as a criterion for predicting academic performance, it appears that all the studies have either represented attendance by a numerical value (i.e. number of attendance occurrences) (Mazza and Milani, 2004), by ordinal values (i.e. poor, average, or high attendance) (Yadav et al., 2012), or by boolean values (i.e. yes/ no attendance) (Quadri and Kalyankar, 2010).

Student transcripts in HE have also always been dealt with in a classical way, i.e. educators rely on building their knowledge about student's performance on their marks as they stand, without any preparations. Additionally, Osmanbegović and Suljić (2012) applied three supervised DM algorithms on the pre-operative assessment data (that included students' GPA as a variable) to predict success in a course (either pass or fail). The data collected from the surveys was gathered during the summer semester at the University of Tuzla, the Faculty of Economics, in the academic year 2010-2011, among first year students and the data was taken during enrolment.

On other hand, Yadav et al. (2012) used Class Test Grade (CTG) as a variable, which is the result of calculating the average from class tests per semester. The authors split the CTG into 3 classes: Poor (less than 40%), Average (greater than or equal to 40%, and less than 60%), and Good (greater than or equal to 60%). Besides other variables, authors used CTG to compare three different DM Decision Tree algorithms. A similar method was used in Baradwaj and Pal (2011) in order to help identify the drop-outs and students who need special attention, to allow the teacher to provide appropriate advice.

Many other studies have used student marks (in its different forms) as an input variable for the EDM process since the beginning of using DM on education data, despite the changes to students' assessment methods (Heywood, J., 2000).

Many researchers concentrated on studying college students' performance, depending on factors such as marks and attendance, without any preparations (Gabrilson, 2008; Luan, 2002; Minaei et al., 2004). However, this research has sought to address this issue of preparation, utilizing the students' academic settings, formulating them, and finally predicting the students' performance by building a prediction model that can provide higher Classification Accuracy (CA) and a much more reliable picture of students' performance in HE.

In the next section, all results of the used DM techniques will be compared and evaluated with other related work. The main issues that will be addressed are: the type and size of used data, the objectives, the methodologies and the DM techniques used.

5.5.1. Random Forest (RF): Comparisons

Some studies have used RF to predict students' performance in HE, but they were using the data as it is or preparing it in a traditional way, which this research has challenged and tried to change. A well-known and recent study that uses RF, is the one by Asif, Hina and Haque (2017). The aim of this study was to use DM techniques for predicting the students' graduation performance in their final year at university, using only pre-university marks (High School Certificate) and examination marks from their early years at university. The data from two academic batches (Civil Engineering Department), included 214 undergraduate students enrolled in the academic years 2005–06 and 2006–07. The maximum CA that they had was 69%, but they were using the data as it was, without any preparation.

Tarekegn and Sreenivasarao (2016), in their study, concentrated on predicting students' placement in HE. In terms of the data that they used, data was taken from the Assistance Registrar Office of Gondar University. The data set contains about 1496 instances of placed students. The data of faculties selected for this study are: Faculty of Natural and Computational Science, Faculty of Agriculture, and Faculty of Engineering. However, they took the student data and applied them to the DM techniques with "cleaning" only and without having any preparation, especially for the transcript data. The maximum CA that they had was 86.56%, using the RF technique.

When this method of applying the data to the DM techniques was tested, the CA value was even lower and the RMSE was high. Table 5.1 shows the

results in terms of data preparation using SAC and CAR, and without SAC and CAR. The results show that when SAC and CAR was used, the CA was higher and the RMSE was lower, compared with not using SAC and CAR (which is the same for other related work).

5.5.2. Artificial Neural Networks (ANN): Comparisons:

There are some studies that have used ANN to predict students' performance in HE but they were also using the data as it is, with cleaning only and in a traditional way. The study by Agrawal and Pandya (2016), aims to investigate if Neural Networks are a fitting classifier to predict student performance from Learning Management System data, in the context of EDM. The dataset used for this study is a Moodle log file containing log information of about 4601 students over 17 undergraduate courses, between 2014 and 2015. They used 13 features in their study but three of them related to the students' transcripts, such as exam marks, assignment marks and quiz marks. However, their preparation was only by normalizing their data.

In Sebastian and Puthiyidam's study (2015), the main objectives was, firstly, to determine all the personal and academic factors that affect the performance of the student, and secondly to transform these factors to a suitable form for system coding. They then wanted to model a Neural Network that could predict the performance, based on the data of student. In their paper, the data was collected from two classes: Eight and Nine of Sacred Hearts Girls' High School, Bharananganam, Kerala. A dataset of 300 students was used for the evaluation. The selected attributes, such as average attendance of students

(good, average) and average marks of students (low, good, and average) were applied as they were, without preparation. The maximum CA that they had was 35%, using the RF technique.

Rasheela Asif (2015) used DM methods to study the performance of undergraduate students. Two aspects of students' performance were focused upon: firstly, predicting students' academic achievement at the end of a four-year study programme and secondly, studying typical progressions and combining them with prediction results. Two important groups of students were identified: the low and high achieving students. The results indicate that by focusing on a small number of courses, you can identify indicators of particularly good or poor performance.

The data used in this study comprises students' marks in a 4-year Information Technology Bachelor Degree of a public-sector engineering university in Pakistan. This study employs data from two academic batches, using a sample of 210 undergraduate students who had enrolled in the academic years of 2007/08 and 2008/09. The dataset comprises variables related to students' pre-admission marks and the marks for all the courses that are taught in the four years of the Degree programme. The preparation of the students' marks included the following: the marks were calculated as the sum of 10% of the first year average examination mark, 20% of the second year, 30% of the third year and 40% of the fourth year average examination mark. The interval of the graduation mark is divided into five possible values/categories: A (90%-

100%), B (80%-89%), C (70%-79%), D (60%-69%), and E (50-59%), as these intervals are well understood by teachers and students alike.

When the data was applied to these DM techniques for the purpose of this thesis, the CA value was lower and the RMSE was high. Table 5.2 shows the results in terms of data preparation using SAC and CAR, and without SAC and CAR.

The results show that when SAC and CAR were used, the CA was higher and the RMSE was lower, compared with not using SAC and CAR (which is the same for other related work).

5.5.3. k-Nearest Neighbours (kNN): Comparisons:

There are some related papers that used kNN to predict students' performance in HE, but similarly to other used DM techniques, they have the same way of preparing the data. The most popular study that uses kNN is the study by Karthikeyan and Kavipriya (2017). This paper focused on improving student performance prediction, based on their personal and academic performance characteristics. The dataset was taken from the CMS College of Science and Commerce, Coimbatore. The size of the dataset was 650 undergraduate students enrolled during the year 2015 and the number of attributes was 52. The data contains variables related to students' internal marks, used to select the students' prior entrance, to final examination marks of the courses that are taught in the first and second academic years of their study. The data in this

study was only “cleaned”, without any preparation and the CA using kNN was 80%.

Kumar and Sharma (2016) concentrated on the performance of students in HE but they took the student data and applied it to DM techniques with normal preparation. When their method of applying the data to DM techniques was tested for this research, the CA was lower and the RMSE was high. Table 5.3 shows the results in terms of data preparation, using SAC and CAR, and without SAC and CAR.

The results show that when SAC and CAR were used, the CA was higher and the RMSE was lower, compared with not using SAC and CAR (which is the same for other related work).

5.6. Conclusions:

The results gained by applying the DM techniques (Random Forest, Artificial Neural Networks, k-Nearest Neighbours) on our dataset have been evaluated by measuring the statistical differences between CA and RMSE of all models. Evaluation has been carried out for all results, to compare all results using DM techniques. It is clear that RF has the highest CA and the lowest RMSE and the importance of SAC and CAR in increasing prediction accuracy has been proved in this chapter.

Finally, the results have been compared with previous studies that predicted students’ final marks, based on students’ marks at earlier stages of their study. The comparisons have taken into consideration similar data and attributes,

whilst firstly excluding average CAR and SAC and secondly, by including them, and then measuring the prediction accuracy between both. The aim of this comparison is to ensure that the new preparation process will positively affect the final results. In chapter 5, the sixth objective has been achieved successfully by evaluating and compares the results of applying different DM techniques, in order to find the optimum technique to predict students' academic performance efficiently.

The second and fifth contributions have been achieved successfully by increasing the accuracy and credibility of the data collected from students' academic profiles - such as their attendance and transcripts - by replacing the traditional ways of handling such data and this research also contributes towards predicting students' academic performance, considering more accurate and credible students' academic data, using different DM techniques.

In chapter 6, a conclusion will be drawn, including a discussion of contributions and limitations. Finally, recommendations and future work will be presented.

CHAPTER SIX: CONCLUSION, DISCUSSION, AND FUTURE WORK

Chapter 6 summarizes the thesis, discusses its findings and contributions, points out limitations of the current work, and outlines directions for future research. The chapter is divided into four sections. Section 6.1 is a summary of the thesis; Section 6.2 presents a discussion of the contributions and limitations of the current work; Section 6.3 discusses future work, and Section 6.4 brings the thesis to a conclusion.

6.1 Summary of the Thesis:

Predicting student performance has been an important area of investigation since the emergence of Educational Data Mining (EDM), as it can be used to improve the outcomes of HE, particularly since predicting students' future academic performance may help direct students towards their strengths and prevent them from continuing their performance retraction in certain areas. This thesis contributes towards predicting students' academic performance, considering more accurate and credible students' academic data, using different DM techniques.

Chapter 2 discussed and investigated different studies that used DM techniques in education. Different research studies, that employed EDM in analyzing higher educational data in particular, have been reviewed. The history of EDM, starting with the definition and applications, to the use of it in predicting student performance in HE, has been presented in Chapter 2, along

with the methods and techniques that are used by researchers to draw conclusions regarding the relationship between initial data and predicted student performance. In chapter 2, the fourth objective has been achieved successfully by investigating the existing studies regarding students' academic performance, including the research objectives, data source and Data Mining (DM) techniques which have been applied.

Chapter 3 described the methodology used in this thesis. Firstly, the source of the data used and their types were characterized, which consists of two main parts: Student Attendance Data and Student Transcript Data. While the former represents module-oriented data, which reflects the characteristics of student attendance at each module, the latter characterizes students' transcript data, with regards to the enrolled module's assessment methods, students' marks at each level of their study, their module marks, and other transcript-related attributes.

Then, the data preparation process was explained, through which data was cleaned, sorted, filtered, transformed a few times, to suit different DM techniques and different phases of the experiments, and various new attributes were constructed to serve this purpose. Two main attributes were constructed and formulated to differentiate this work from previous research that looked at students' academic performance. The newly formulated attributes are: Student Attendance Credibility (SAC) and Coursework Assessment Ratio (CAR). In chapter 3, the first three objectives have been achieved successfully by investigating the factors of students' academic profiles from the data set used,

identifying those which have the biggest impact on their academic performance, using those factors in a way that guarantees more accuracy and credibility and showing the differences between using data collected from students' academic records as it is, and using formulas that represent these data in a more accurate and credible form.

In Chapter 4, data that was fed into DM techniques had been pre-processed through different stages. The stages included adding new attributes that increased the accuracy of prediction, and other various pre-processing stages to guarantee clean and ready data. After splitting the modified data into 70% for training and 30% for testing, the three DM techniques (Random Forest, Neural Networks, and k-Nearest Neighbors) were trained on the data that contained records of students who completed their entire programme (i.e. including their Average A, B, C, SAC and CAR). After that, experiments of used DM techniques were applied to investigate the prediction of final students' marks of the second year, based on the first year's marks, prediction of final students' marks of the third year, based on the first and second year's marks and prediction of final students' marks of the third year, based on the first year's marks only.

Additionally, in terms of SAC and CAR in predicting students' marks, further investigation was carried out such as, predicting students' average marks while excluding SAC, and in another experiment including it, predicting students' average marks while excluding CAR, and in another experiment including it and predicting students' marks while including both CAR and

SAC. Finally, average SAC and CAR were removed from the dataset, which helped in determining the difference between using the newly formulated attributes (SAC and CAR) and not using them in the prediction task. In chapter 4, the fifth objective has been achieved successfully by showing the importance of applying the DM classification techniques on HE students' data to predict and classify students' academic performance, while taking into consideration the accuracy of input data.

In Chapter 5 the results were generated by applying the DM techniques on our data set and evaluated by measuring the statistical differences between Classification Accuracy (CA) and Root Mean Square Error (RMSE) of all models. Comprehensive evaluation has been carried out for all results in the experiments to compare all DM techniques results, and it has been found that Random forest (RF) has the highest CA and lowest RMSE. The importance of SAC and CAR in increasing the prediction accuracy has been proved in Chapter 5.

Finally, the results have been compared with previous studies that predicted students' final marks, based on students' marks at earlier stages of their study. The comparisons have taken into consideration similar data and attributes, whilst first excluding average CAR and SAC and secondly by including them, and then measuring the prediction accuracy between both. The aim of this comparison is to ensure that the new preparation process stage will positively affect the final results. In chapter 5, the sixth objective has been achieved successfully by evaluating and compares the results of applying different DM

techniques, in order to find the optimum technique to predict students' academic performance efficiently.

6.2 Discussion:

In this section, the discussion will be divided into two sections: the first section will discuss the contributions of the thesis and the second section will discuss the limitations of the current work in detail.

6.2.1 Contributions of the Thesis:

- Modifying the Educational Data Mining (EDM) pre-processing part to improve the quality of the model outcome. In this thesis, the modifications of the pre-processing part of EDM have proved better outcomes using new proposed parameter engineering techniques such as SAC (Section 3.3.1) and CAR (Section 3.3.2).
- There are multiple ways to prepare any data such as cleaning, filtering, categorizing, etc., but in this thesis there is a new way has been used in the preparation phase which is "data filtering" that can be used for any kind of educational data.
- In term of the new parameter engineering techniques for the two datasets, one is Students Attendance Credibility (SAC) and other is Course Assessment Ratio (CAR). SAC is referring to attendance dataset and CAR referring to transcripts dataset.

- This thesis has increased the accuracy and credibility of the data collected from students' academic profiles, such as their attendance and transcripts, by replacing the traditional methods of handling such data.
- Students' transcript data has been formulated in a way that guarantees that the assessment methods of students' enrolled module reflect the original data. In other words, a contribution of this thesis is to consider the differences between various assessment methods and refine student marks based on these differences.
- Student attendance has been formulated in a way that takes into consideration other factors that have an impact on students' attendance (such as number of times the instructors record attendance), rather than using the attendance data as it is (e.g. number of absences, poor or good attendance, etc.).
- Based on the above, this research has also contributed towards predicting students' academic performance, considering more accurate and credible students' academic data, using different DM techniques.

6.2.2 Limitations of the Current Work:

Although the research has achieved its aims, there were some unavoidable limitations that should be discussed. In terms of data sources, the data was collected from the Admission Department of a UK university, which is the typical way of collecting the data.

In terms of the related work, all the literature uses data collected from students' academic records as it stands and does not use formulas that

represent more accurate and credible forms of data. However, it was very difficult to find any related work on this subject.

In terms of data variety, there are good amounts of data related to the academic performance of the student, but there are few that consider other aspects of university life. The model could benefit from integration with other data sources with different background data, information of the family, academic and socio-economic. As well as the use of new data sources, it is important to consider new ways of gathering information from the students and their interactions, taking advantage of social media and other communication tools.

6.3 Future Work:

This thesis focuses on Educational Data Mining (EDM) research at a UK university and can be developed in various ways. However, using different DM techniques to learn the classification models, or a combination of classifiers, to improve the performance results. This work has been applied on one university and for future work it could be applied on different universities inside and outside the UK, to look at data with different assessment methods.

The research could also use data from different programs as this thesis only used data from a Computer Science program for the model. The inclusion of other programs in the model could bring a different perspective and allow the university to gain a better understanding of dropout behaviour and patterns.

This thesis focused on students' academic performance based on students' attendance and transcript data. However, there are many factors which could impact on academic performance, such as the time of class during the semester. Moreover, finding the correlation between the modules to decide which one should be chosen for the student depends on his/her performance could assess the module for any program.

In term of using data mining techniques, Random Forest (RF), Artificial Neural Network (ANN) and k-Nearest Neighbours (kNN) gave better results for the used data in this thesis, for future work, using different dataset such as graduate students data from different programs may need applying different DM techniques to give the best results depends on the used data. The same dataset could be used but with different aims objectives such as predicting the future career of student.

For future research, data would be collected from different sources such as the Learning Management System (LMS), Virtual Learning Environment (VLE), Mobile applications, Wikis, blogs, and forums, in order to have data with low missing values. In addition, data in this research have been applied for one department (computer science), for future work, data would be applied for different departments.

Finally, predicting the performance of module it self rather than student's performance by comparing and evaluating these modules using parameter engineering technique.

REFERENCES

- Alsuwaiket, M., Dawson, C.W. and Batmaz, F., (2016). Measuring the credibility of student attendance data in Higher Education for data mining. Presented at the 5th International Conference on Knowledge and Education Technology, University of Hertfordshire, October 29-31st.
- Avouris, N., Komis, V., Fiotakis, G., Margaritis, M., Voyiatzaki, E., (2005). Why logging of fingertip actions is not enough for analysis of learning activities. In: Workshop on Usage Analysis in Learning Systems, pp. 1–8. AIED Conference, Amsterdam.
- Ayers E., Junker B.W. (2006). Do skills combine additively to predict task difficulty in eighth grade mathematics? In AAAI Workshop on Educational Data Mining: Menlo Park, pp. 14-20.
- Al-Radaideh Q., Ahmad Al-Ananbeh, Emad. M. Al-Shawakfa, (2011) A Classification Model for Predicting the Suitable Study Track for School Students, International Journal of Research and Reviews in Applied Sciences, 8(2): pp. 247-252.
- Al-Radaideh Q., and Eman A., (2012), Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher, 3(2): pp. 144-151.
- Andrew. P. Bradley. (1997). The Use of the area under the ROC Carve in the evaluation of machine learning algorithms". Pattern Recognition, Vol. 30, No. 7, pp. 1145-1159.

- Anna Lukkarinen, Paula Koivukangas, Tomi Seppälä. (2016). Relationship between Class Attendance and Student Performance. Procedia - Social and Behavioral Sciences. Volume 228, 20 July 2016, Pages 341-347.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), pp. 5-32.
- Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.
- Brijesh K. & Saurabh P. (2011), "Mining Educational Data to Analyze Students‟ Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.
- Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended programming tasks. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.
- Breiman L, Freidman J., Olshen R. and Stone C. (1984). Classification and Decision Trees, Wadsworth.
- Brusilovsky, P., Peylo, C. (2003). Adaptive and intelligent web-based educational systems. Int. J. Artif. Intell. Educ. 13(2–4), pp. 159–172.
- Brusilovsky, P., Miller, P. Leggett. (1999). Web-based testing for distance education. In: De Bra, WebNet'99, World Conference of the WWW and Internet, pp. 149–154. AACE, Honolulu.
- Brijesh K., Saurabh P., (2011). Mining Educational Data to Analyze Students' Performance, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6.

- Burlak, G., Muñoz, J., Ochoa, A., Hernández, J.A. (2006). Detecting Cheats In Online Student Assessments Using Data Mining. In International Conference on Data Mining, Las Vegas, pp. 204-210.
- Bharadwaj B.K. and Pal. S. (2011). Mining Educational Data to Analyze Students' Performance, International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.
- Bruha, I ., Famili, A. (2000). Post-processing in machine learning and data mining, in ACM SIGKDD Explorations Newsletter Volume 2, Issue 2.
- Behan, Donald F. (2009). Statistical Credibility Theory, South-eastern Actuarial Conference.
- Chan, C.C. (2007). A Framework for Assessing Usage of Web-Based e-Learning Systems. In International Conference on innovative Computing, Information and Control, Washington, DC, pp. 147- 151.
- Cristobal Romero and Sebastian Ventura (2013). Data Mining in Education. WIREs Data Mining Knowledge Discovery, 3: pp. 12–27 doi: 10.1002/widm.1075.
- Cristóbal R., José Raúl R. and Sebastián V. (2014). A Survey on Pre-Processing Educational Data, Studies in Computational Intelligence 524, DOI: 10.1007/978-3-319-02738-8_2, Springer International Publishing Switzerland.
- Cristóbal R., Sebastian V. (2010). Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) 40(6): pp. 601 – 618.
- Chen, c., Chen, M., Li, Y. (2007). Mining key formative assessment rules

based on learner portfiles for web-based learning systems. In IEEE International Conference on Advanced Learning Technologies, Japan, pp. 1- 5.

- Chiu, D.Y., Pan, Y.C., Chang, W.C. (2008). Using rough set theory to construct e-learning faq retrieval infrastructure. In IEEE Ubi-Media Computing Conference. Lanzhou, pp. 547-552.
- Cox, K., Eick, S., and Wills, G. (1997). Visual data mining: recognizing telephone-calling fraud. Data Mining and Knowledge Discovery.
- Dringus, L., & Ellis, T. (2005). Using data mining as a strategy for assessing asynchronous discussion forums. Computers & Education, 45, pp. 141-160.
- David R. (1993). Do Students Go to Class? Should They? The Journal of Economic Perspectives (summer): pp. 167-174.
- Delavari, N., Phon-Amnuaisuk, S., Beikzadeh, M. (2008). Data mining application in higher learning institutions. Inf. Educ. J. 7(1), pp. 31–54.
- Delgado, M., Gibaja, E., Pegalajar, M.C., Pérez, O. (2006). Predicting Students' Marks from. Moodle Logs using Neural Network Models. In International Conference on Current Developments in Technology-Assisted Education, Sevilla, Spain, pp. 586-590.
- Edward G. Carmines, Richard A. Zeller. (1987). Reliability and Validity Assessment.
- Fayyad, U. M. et al. (1996). From data mining to knowledge discovery: an overview. In Fayyad, U. M.et al (Eds.). Advances in knowledge discovery and data mining. AAAI Press / The MIT Press.

- Grubbs F. E. (2011). Procedures for detecting outlying observations in samples. pp.1–21, Technometrics.
- Fausett, L.V., Elwasif, W. (1994). Predicting performance from test scores using backpropagation and counterpropagation. In IEEE World Congress on Computational Intelligence, Paris, France, pp. 3398–3402.
- Gabrilson, S., Fabro, D. D. M., Valduriez, P., (2008). Towards the efficient development of model transformations use model weaving and matching transformations, Software and Systems Modeling. Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education.
- Getaneh B. T., Vuda S. (2016). Application of Data Mining Techniques to Predict Students Placement in to Departments. International Journal of Research Studies in Computer Science and Engineering. Volume 3, Issue 2, pp. 10-14.
- George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.
- Graham G. (1999). Using Assessment Strategically to Change the Way Students Learn, From Assessment Matters in Higher Education, ISBN 0749411139.
- Gedeon, T.D., Turner, H.S. (1993). Explaining student grades predicted by a neural network. In International conference on Neural Networks, Nagoya, pp. 609-612.
- Graham G. and Claire S. (2004). Conditions Under Which Assessment Supports Students' Learning", Journal of Learning and Teaching in Higher

Education, Issue 1.

- Gonzalo M. and Oscar M. (2010). A survey of data mining and knowledge discovery process models and methodologies, The Knowledge Engineering Review, Volume 25, Issue 2 June 2010, pp. 137-166.
- Haddawy, P., Thi, N., Hien, T.N. (2007). A decision support system for evaluating international student applications. In Frontiers In Education Conference, Milwaukee, pp. 1-4.
- Hämäläinen, W., Vinni, M. (2006). Comparison of machine learning methods for intelligent tutoring systems. In international conference in intelligent tutoring systems, Taiwan, pp. 525-534.
- Han, J., Kamber, M. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco.
- Heywood, J. (2000). Assessment in Higher Education: Student learning, teaching programmes and institutions. London: Jessica Kingsley.
- Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, 61, pp.133-145.
- Hughes, G. & Dobbins, C. RPTEL (2015). The utilization of data analysis techniques in predicting student performance in massive open online courses (MOOCs). Research and Practice in Technology Enhanced Learning. pp. 1793-7078.
- Hien, N.T.N., Haddawy, P. (2007). A decision support system for evaluating international student applications. In Frontiers In Education Conference, Milwaukee, pp. 1-6.

- Romesburg H.C. (2004). Cluster Analysis for Researchers. Morrisville, NC: Lulu.com. (Reprint of 1984 edition, with minor revisions.)
- Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications, 3(5), pp. 605-610.
- I.H. Witten and E. Frank. (2012). Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)
- Joppe, M. (2000). The Research Process. Retrieved February 25, 2015, from http://www.ryerson.ca/~mjoppe/rp.htm.
- John T.E. Richardson. (2015). Coursework versus examinations in end-of-module assessment: a literature review, Assessment & Evaluation in Higher Education, Volume 40, 2015 - Issue 3.
- Joseph M. Hellerstein. (2008). Quantitative Data Cleaning for Large Databases.
- Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowledge-Based Systems, 23(6), pp. 529-535.
- Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques, In International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Oxford, pp. 3-5.

- Karthikeyan K. and Kavipriya P. (2017). On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5.
- Knapp, T. R. (1991). Coefficient alpha: Conceptualizations and anomalies. Research in Nursing & Health, 14, pp. 457-480.
- Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. In Computer & Education Journal, 53,3, pp. 950-965.
- Liu, H., Motoda, H. (2007). Computational Methods of Feature Selection. Chapman & Hall/CRC, Boca Raton.
- Lehmann, E. L.; Casella, George (1998). Theory of Point Estimation (2nd ed.). New York: Springer. ISBN 0-387-98502
- Mikut, R., & Reischl, M. (2011). Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(5), pp. 431-443.
- Morris, P.E., & Fritz, C.O. (2015). Conscientiousness and procrastination predict academic coursework marks rather than examination performance, Learning and Individual Differences.
- Mostow J. and Beck J., (2006). Some useful tactics to modify, map and mine data from intelligent tutors. Natural Language Engineering 12(2), pp. 195-208.
- Minaei-Bidgoli B., Kortemeyer G., Punch W.F. (2004). Enhancing Online

Learning Performance: An Application of Data Mining Methods, In Proceeding of Computers and Advanced Technology in Education.

- Mohammed J. Zaki and Wagner Meira Jr. (2013). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press.
- Mor, E., Minguillón, J. (2004). E-learning personalization based on itineraries and long-term navigational behavior. In: Thirteenth World Wide Web Conference, pp. 264–265. ACM, New York.
- Márquez-Vera, C., Cano, A., Romero, C., Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Appl. Intell. 38(3), pp. 315–330.
- Monika G. and Rajan V. (2012). Applications of Data Mining in Higher Education, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1.
- Mazza, R. (2009). Introduction to Information Visualization. Springer.
- Martinez, D., (2001). Predicting Student Outcomes Using Discriminant Function Analysis. In Meeting of the Research and Planning Group, Lake Arrowhead, CA, pp. 1-22.
- Minaei-bidgoli B., Kashy D.A., Kortmeyer G., Punch W.F. (2003). Predicting student performance: an application of data mining methods with an educational Web-based system. In International Conference on Frontiers in Education, pp. 13-18.
- Njål Foldnes (2017). The impact of class attendance on student learning in

a flipped classroom. Nordic Journal of Digital Literacy. 01-02 / 2017 (Volum 12)

- Quadri M. N. and Kalyankar N.V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. Global Journal of Computer Science and Technology, Vol. 10 Issue 2 (Ver 1.0).
- Nebot, A., Castro, F., Vellido, A., Mugica, F. (2006). Identification of fuzzy models to predict students perfornance in an e-learning environment. In International Conference on Web-based Education, Puerto Vallarta, pp. 74-79.
- Nilakant, K., Mitrovic, A. (2005). Application of data mining in constraint based intelligent tutoring systems. In: International Conference on Artificial Intelligence in Education, pp. 896–898. Amsterdam.
- Upadhyay N., Katiyar V. (2014). A Survey on the Classification Techniques in Educational Data Mining, International Journal of Database Management System (IJDMS), 3(11), pp. 725–728.
- Osmanbegović, E., & Suljić, M. (2012). Data mining approach for predicting student performance. Economic Review, 10(1).
- Ajith, P., Tejaswi B. (2013). "Rule Mining Framework for Students Performance Evaluation". International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6.
- Paul Bridges, Angela Cooper, Peter Evanson, Chris Haines, Don Jenkins, David Scurry, Harvey Woolf & Mantz Yorke. (2002). Coursework Marks High, Examination Marks Low: Discuss, Assessment & Evaluation in Higher Education. Volume 27, Issue 1.

- Pritchard, D., Warnakulasooriya, R. (2005). Data from a Web-based Homework Tutor can predict Student's Final Exam Score. In World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake, pp. 2523-2529.
- Robin N. and Jeremy S. (2005). Schooling effects on subsequent university performance: evidence for the UK university population, 2005, retrieved form http://dx.doi.org/10.1016/j.econedurev.2004.07.016 .
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), pp. 601-618.
- Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers & Education, 68, pp. 458-472.
- Duda R.O. and Hart P.E. (1973). Pattern Classification and Scene Analysis, New York: John Wiley & Sons, the practical applications of knowledge discovery and data mining, pp. 29-39.
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), pp. 601-618.
- Ross J. Quinlan. (1992). Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348.
- Romero, C. and Ventura, S. (2007). Educational Data mining: A survey from 1995 to 2005, Expert systems With Application. pp. 135-146.

- Ross J. Quinlan. (1992). Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343-348.
- Romero C, Ventura S, Pechenizky M, Baker R. (2010). Handbook of Educational Data Mining. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press.
- Kohavi R. and Provost F. (1998). Glossary of Terms, in Spec. Issue on Apps of Machine Learning and the KDD Process, Machine Learning Journal, 30, pp. 271-274. Kluwer.
- Romero, c., ventura, s., hervás, c., gonzales, p. (2008). Data mining algorithms to classify students. In International Conference on Educational Data Mining, Montreal, Canada, pp. 8-17.
- Riccardo Mazza and Christian Milani. (2004). GISMO: a Graphical Interactive Student Monitoring Tool for Course Management Systems", T.E.L.'04 Technology Enhanced Learning '04 International Conference. Milan, pp. 18-19.
- Raj K. and Akshita S. (2016). Comparative analysis of SVM and kNN for academic prediction of students. IJITKM. Volume 10, Number 1.
- Rasheela A. (2015). Predicting Student Academic Performance at Degree Year: A Case Study. I.J. Intelligent Systems and Applications, 01, 49-61.
- Raheela A., Saman H., Saba I. (2017). Predicting Student Academic Performance using Data Mining Methods. International Journal of Computer Science and Network Security, Volume.17, No.5.

- Kotsiantis S. B. Zaharakis I. Pintelas D. P. E.(2007). Machine learning: a review of classification and combining techniques. Artif Intell Rev (2006) 26: pp. 159–190.
- Suthers, D., Verbert, K., Duval, E., Ochoa, X. (eds.), Clow, D. (2013). MOOCs and the funnel of participation. International Conference on Learning Analytics and Knowledge, pp. 185–189. ACM New York, NY.
- Salmeron-Majadas, S., Santos, O., Boticario, J.G., Cabestrero, R., Quiros, P. (2013). Gathering emotional data from multiple sources. In: D'Mello, S.K., Calvo, R.A., Olney, A. (eds.) 6th International Conference on Educational Data Mining, pp. 404–405. International Educational Data Mining Society, Memphis.
- Shangping, D., Ping, Z. (2008). A data mining algorithm in distance learning. In International Conference on Computer Supported Cooperative Work in Design, Xian, pp. 1014-1017.
- Hijazi S. T., and Naqvi R. S. (2006) Factors affecting student's performance: A Case of Private Colleges, Bangladesh e-Journal of Sociology, Vol. 3, No. 1.
- Yadav S. K. et al. (2012). Data Mining Applications. A comparative Study for Predicting Student's performance, International Journal of Innovative Technology & Creative Engineering (ISSN: 2045-711) VOL.1 NO.12.
- Sumam S. and Jiby J P. (2015). Evaluating Students Performance by Artificial Neural Network using WEKA. International Journal of Computer Applications (0975-8887), Volume 119- No.23.
- Tang, C., Lau, R.W.H., Li., Q., Yin, H., Li, T., Kilis, D. (2000).

Personalized courseware construction based on web data mining. In First International Conference on Web Information Systems Engineering, Hong Kong, China, pp. 204-211.

- Ueno M. (2004). Online outlier detection system for learning time data in e-learning and its evaluation. In: International Conference on Computers and Advanced Technology in Education. Beijiing, China, pp. 248– 253
- Vee, M.N., Meyer, b., Mannock, K.L. (2006). Understanding novice errors and error paths in Object-oriented programming through log analysis. In Workshop on Educational Data Mining, Taiwan, pp. 13-20.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference
- Wettschereck, D.: Educational data pre-processing. In: ECML'02 Discovery Challenge Workshop, pp. 1–6. University of Helsinki, Helsinki (2002).
- Wu, A., Leung, C. (2002). Evaluating learning behavior of Web-Based Training (WBT) using Web log. In International Conference on Computers in Education, New Zealand, pp. 736- 737.
- Want, T., Mitrovic, A. (2002). Using Neural Networks to Predict Student's Performance. In International Conference on Computers in Education, Washington, DC, pp. 1-5.
- Wahbeh A., Al-Radaideh Q., Al-Kabi M., and Al-Shawakfa E., (2011), A Comparison Study between Data Mining Tools over some Classification Methods, International Journal of Advanced Computer Science and Applications (IJACSA) - A Special Issue on Artificial Intelligence, SAI

Publisher, 2(0): pp. 18-26.

- Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. arXiv preprint arXiv:1202.4815.
- Yu, C.H., Jannasch-pennell, A., Digangi, S., Wasson, B. (1999). Using On-line interactive statistics for evaluating Web-based instruction, In Journal of Educational Media International, 35, pp. 157-161.
- Zhu, F., Ip, H., Fok, A., Cao, J. (2007). PeRES: A Personalized Recommendation Education System Based on Multi-Agents & SCORM. In: Leung, H., Li, F., Lau, R., Li, Q. (eds.) Advances in Web Based Learning—ICWL 2007. LNCS, vol. 4823, pp. 31–42. Springer, Heidelberg.

BIBLIOGRAPHY

- ALSUWAIKET, M., DAWSON, C.W. and BATMAZ, F., (2018). Measuring the credibility of student attendance data in Higher Education for data mining. International Journal of Information and Education Technology, 8(2), pp. 121-127.
- M. Hagan et al. (2017). Neural Network Design, 2nd Edition, ebook, retrieved from http://hagan.okstate.edu/NNDesign.pdf#page=469 on 18 May 2017.
- Release, M. A. T. L. A. B. (2013). The MathWorks. Inc., Natick, Massachusetts, United States, 488.
- Richa A., Mitula H. (2016). Data Mining With Neural Networks to Predict Students Academic Achievements. IJCST Vol. 7, ISSue 2.
- Wirth, Rüdiger, and Jochen Hipp. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.

APPENDIX A: FIRST PAPER

Citation: DAWSON, C.W., ALSUWAIKET, M. and BATMAZ, F., 2018. Measuring the credibility of student attendance data in Higher Education for data mining. International Journal of Information and Education Technology, 8(2), pp. 121-127.

No. of citations:5

Additional Information:

This paper was accepted for publication in the journal International Journal of Information and Education Technology and the definitive published version is available at:

http://www.ijiet.org/show-97-1170-1.html

This Paper was presented at the *5th International Conference on Knowledge and Education Technology (ICKET 2016) and subsequently published in the Inter- national Journal of Information and Education Technology.*

Publisher: IJIET

Measuring the Credibility of Student Attendance Data in Higher Education for Data Mining

Mohammed Alswaiket, Dr. Christian Dawson, and Dr. Firat Batmaz

Abstract:

Educational Data Mining (EDM) is a developing discipline, concerned with expanding the classical Data Mining (DM) methods and developing new methods for discovering the data that originate from educational systems. It aims to use those methods to achieve a logical understanding of students, and the educational environment they should have for better learning.

These data are characterized by their large size and randomness and this can make it difficult for educators to extract knowledge from these data. Additionally, knowledge extracted from data by means of counting the occurrence of certain events is not always reliable, since the counting process sometimes does not take into consideration other factors and parameters that could affect the extracted knowledge.

As a case example of the above problem, student attendance in Higher Education has always been dealt with in a classical way, i.e. educators rely on counting the occurrence of attendance or absence building their knowledge about students as well as modules based on this count. This method is neither credible nor does it necessarily provide a real indication of a student's performance.

This study explores the above problem and tries to formulate the extracted knowledge in a way that guarantees achieving accurate and credible results. Student attendance data, gathered from the educational system, were first cleaned in order to remove any randomness and noise, then various attributes were studied so as to highlight the most significant ones that affect the real attendance of students. The next step was to derive an equation that measures the Student Attendance's Credibility (SAC) considering the attributes chosen in the previous step. The reliability of the newly developed measure was then evaluated in order to examine its consistency. Finally, the J48 DM classification technique was utilized in order to classify modules based on the strength of their SAC values.

Results of this study were promising, and credibility values achieved using the newly derived formula gave accurate, credible, and real indicators of student attendance, as well as accurate classification of modules based on the credibility of student attendance on those modules.

Index Terms— EDM, Credibility, reliability, Student attendance, Higher Education.

APPENDIX B: SECOND PAPER

Submitted on 10/10/2017 (Under Review)

Computers & Education: An International Journal

https://ees.elsevier.com/cae/default.asp

Journal Metrics

CiteScore: 5.50

More about CiteScore

Impact Factor: 3.819

5-Year Impact Factor: 5.047

Source Normalized Impact per Paper (SNIP): 3.193

SCImago Journal Rank (SJR): 2.613

Formulating Module Assessment for Improved Academic Performance Predictability In Higher Education

Mohammed Alswaiket, Christian Dawson, and Firat Batmaz

Abstract:

The choice of an effective student assessment method is an issue of interest in Higher Education. Various studies (Romero, et al., 2010) have shown that students tend to get higher marks when assessed through coursework-based assessment methods - which include either modules that are fully assessed

through coursework or a mixture of coursework and examinations – than assessed by examination alone. There are a large number of Educational Data Mining (EDM) studies that pre-processed data through the conventional Data Mining processes including the data preparation process, but they are using transcript data as it stands without looking at examination and coursework results weighting which could affect prediction accuracy. This paper proposes a different data preparation process through investigating more than 230,000 student records in order to prepare students' marks based on the assessment methods of enrolled modules. The data have been processed through different stages in order to extract a categorical factor through which students' module marks are refined during the data preparation process. The results of this work show that students' final marks should not be isolated from the nature of the enrolled module's assessment methods; rather they must be investigated thoroughly and considered during EDM's data pre-processing phases. More generally, it is concluded that Educational Data should not be prepared in the same way as exist data due to the differences such as sources of data, applications, and types of errors in them. Therefore, an attribute, Coursework Assessment Ratio (CAR), is proposed to use in order to take the different modules' assessment methods into account while preparing student transcript data. The effect of CAR on prediction process using Random Forest classification technique has been investigated. It is shown that considering CAR as an attribute increases the accuracy of predicting students' second year averages based on their first year results.

APPENDIX C: SUMMARY OF RELATED WORK

The following table shows a summary of related work, the attributes have the author name/s, reference title, study objectives, data source, used data mining techniques, study level and the results of studies. The purpose of this summary is to show the most popular and important related work in term of students' academic performance in Higher Education, the categorizations have shown four important issues, first is presenting the objectives of the studies, second is focusing on the used data sources, third is showing the used data mining techniques and finally the results of these studies have been presented in detail.

Author Name/s	Ref. Title	Study Objectives	Data Source	Used Techniques	Students Level	Study Results
GareyDurden	Garey DurdenC.and Larry V. Ellis, (1995). The Effects of Attendance on Student Learning in Principles of Economics. The American	Concerning with the performance and GPA of students in Higher Education . The author looked at his research at attendance rates of the students and they found that a lower attendance rate negatively affects a student's GPA.	Data was collected from students' attendance sheet in the school	Classification Techniques (Decision tree)	Higher Education	Authors had been found the lower number of students attendance rate negatively affects a student's GPA

	Economic Review (May):343-346.					
David Romero	David Romer, (1993). Do Students Go to Class? Should They? The Journal of Economic Perspectives (summer):167-174.	Concerning with the performance and GPA of students in Higher Education . The author looked at this research at attendance rates of the students and they found that a lower attendance rate negatively affects a student's GPA.	Data was collected from students' attendance sheet in the school	Classification Techniques (Decision tree and naïve bayes)	High er Educ ation	Authors had been found the lower number of students attendance rate negatively affects a student's GPA
Amy Wolaver	Amy M. Wolave, (2002). Effects of Heavy Drinking in College on Study Effort, Grade Point Average, and Major Choice. Contemporary Economic Policy (October):415-428.	She focused in her study on student GPA especially students' alcohol consumption,	Students Questionnaire and "National statistics"	Classification techniques, Clustering, Decision tree	Higher Education	Author had been found that as students drink more, there is a negative affect increment on their GPAs.
Gabrilson	Gabrilson, S., Fabro, D. D. M., Valduriez, P., (2008). Towards the efficient development of model transformations use model weaving	He finds the effective factors to determine a student's test grade, and then the author adjusts these factors to improve the performance of	Data was collected from the results of testes of the students	DM prediction technique (classification, Predictive model, and Bayesian classification)	Higher Education	the author adjusts some factors that affect the students' test grades to

	and matching transformations, Software and Systems Modeling. Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education	students' test grade.				improve the performance of students' test grade.
Luan	Luan, J., (2002). Data mining and knowledge management in Higher Education – potential applications. In Proceedings of AIR Forum, Toronto, Canada	She groups students and determines which students can easily pile up their courses and take courses for longer period of time.	Data was collected from students records	Clustering technique	Higher Education	Author had been helped the universities to identify the requirements of each group of student and then make the best decisions on how to offer courses and curriculum
Minaei-Bidgoli et al.	Minaei-Bidgoli, B., Kortemeyer, G., Punch,W.F., (2004). Enhancing Online Learning Performance: An Application of Data Mining Methods, In Proceeding of	Authors used DM classification techniques to predict students final grade based on their web-use feature. By this technique they can identify students at risk early and	Data was collected from the university by collecting the information of students	DM classification techniques	Higher Education	By using the final results from predicting the final grades based on students' web-use feature; the teacher be able to provide good advice in a timely manner.

	Computers and Advanced Technology in Education.	allow the teacher to provide good advice in a timely manner.				
Shaeela et al	Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, (2010). "Data Mining Model for Higher Education System", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29.	Predict the activities of student's learning.	Data was collected from the students' activities in learning.	k-means clustering algorithm	Higher Education	Authors had been found that the information generated after the implementation phase of data mining technique was helpful for both teacher and students
Zekić-Sušac, Frajman-Jakšić and Drvenkar	Zekić-Sušac M., Frajman-Jakšić A. &Drvenkar N. (2009), Neuron Networks and Trees of Decision-making for Prediction of Efficiency in Studies, Vol.No.2, pp. 314-327.	The authors predicted college students' performance by creating a model using artificial neural networks (ANN) and classification trees decision-making, and with the analysis of factors which influence college students' success in Higher	Data was collected students records by collecting their grades (GPA)	Artificial neural networks (ANN)	Higher Education	By using ANN and classification trees decision-making the authors increased the accuracy of success of students to increase their performance

		Education .				
Khan, Z. N.	Khan, Z. N., (2005). "Scholastic Achievement of Higher Secondary Students in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87.	In his study conducted a performance study on 400 students (200 male and 200 female) to establish the prognostic value of different measures of (cognition, personality and demographic variables) for success at high school level in science. The selection in this study was based on cluster technique in which the entire population was divided into groups (clusters).	Data was collected from 400 students (200 male, 200 female) by looking on their courses and their grades on each course	Clustering	High School	Author had been found that the females had higher academic achievement in science and the males had low academic achievement in general.
Beikzadeh,M and Delavari N	Blikstein, P. (2011). Using learning analytics to assess students' behavior in open-ended	the authors applied the data mining (classification and predictive techniques) in educational area to	Data was collected students records by collecting their grades (GPA)	Classification and predictive techniques (Decision tree and naïve bayes)	Higher Education	Results of this research is adding some values of enhancement the education process so by using the data mining they

	programming tasks. Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta, Canada.	identify and improve educational process in Higher Education al system, which can enhance their decision making process				enhanced the decision making process
Yu et al	Yu, C. H., DiGangi, S., Jannasch-Pennell, A., Lo, W., &Kaprolet, C. (2007). A data-mining approach to differentiate predictors of retention. In the Proceedings of the Educause Southwest Conference, Austin, Texas, USA.	Applied DM techniques to differentiate the predictors of retention for freshmen enrolled at Arizona State University. They used classification decision tree based on entropy tree-splitting.	Data was collected from freshmen students who enrolled at Arizona State University. The data was the retention for freshmen students	Classification decision tree based on entropy tree-splitting	Higher Education	Authors had been found that the cumulated earned hours were the most significant factor contributing to retention
1. S. Kotsiantis, Pierrakeas&Pintelas	S. Kotsiantis, C. Pierrakeas, P. Pintelas (2004). Predicting Students Performance in Distance Learning	Their goal was to predict students' performance. The authors used two different ways for prediction:	National statistics data by using some information about the students like (grade, courses, final GPA)	used in their study key demographic variables and assignment marks in some algorithms of	Higher Education	The authors achieved their main goal by increasing the accuracy of success, the advantages of the research were using

	Using Machine Learning Techniques, Applied Artificial Intelligence (AAI), Volume 18, Number 5.	1. Demographic variables only, accuracy varied from 58.84% "when using neural network" to 64.47% "when using support vector machines". 2. Other variables beside demographic, naïve Bayes classifier was the most accurate algorithm for predicting students' performance.		supervized machine learning like (logistic regression, decision trees, naïve Bayes classifier, instance-based learning, support vector machines and artificial neural networks "ANN")		more than classification technique, so the study considered as a comprehensive study to predict the students' performance and they found after comparing the results that naïve bayes was the most accurate algorithm, also by using ANN, they increased the accuracy from 58.84% to 64.47%.
P. Ajith, B. Tejaswi, M.S.S.Sai	P. Ajith, B. Tejaswi, M.S.S.Sai (2013). "Rule Mining Framework for Students Performance Evaluation". International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013	The main objective in this study is to help the students and the teachers to improve the division of the student and identify those students which needed special attention to reduce fail percentages and taking appropriate action for the upcoming	The data set used in this study was obtained from B.TECH students of KITS engineering college from session 2006 to 2010. Initially size of the data is 30.	Authors had been used the classification task on student database to predict the students division on the basis of previous database. As there are many approaches that are used	Higher Education	A result of a tree based classification was complicated to understand and depends on the technical competency of the decision maker. Among sets of items in transaction databases, Association Rules aims at discovering implicative

		evaluations.		for data classification, the decision tree method is used in this paper.		tendencies that can be valuable information for the decision-maker which is absent in tree based classifications. Compared to tree based classifications the results are better to understand and can be applied to real time use.
Agathe Merceron and Kalina Yacef	Agathe Merceron and Kalina Yacef, (2005). "Educational Data Mining: a Case Study". proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press.	The purpose of this study is to synthesize and share the authors' various experiences of using Data Mining for Education system, especially to support reflection on teaching and learning, and to contribute to the emergence of stereotypical directions.	Authors had been performed a number of queries on datasets collected by the Logic-ITA to assist teaching and learning. The Logic-ITA is a web-based tutoring tool used at Sydney University since 2001, in a course taught by the second author.	Authors had been worked with Excel and Access to perform simple SQL queries and visualization, and then they used Clementine for clustering, Tada-Ed for clustering, classification and association rule, also they used SODAS to perform	Higher Education	support learner reflection and provide proactive feedback to learners.

				symbolic data analysis.		
2. El-Halees, A.	El-Halees, A. (2008) "Mining Students Data to Analyze Learning Behavior: A Case Study". The 2008 international Arab Conference of Information Technology (ACIT2008) –Conference Proceedings, University of Sfax, Tunisia.	In this paper the author had been used educational data mining to analyze learning behavior	In this paper the author had been collected the students' data from database management system course held at the Islamic University of Gaza in the first semester of 2007/2008. The number of students was 151. The sources of collected data were: personal records and academic records of students, course records and data came from e-learning system. For e-learning system the course used Moodle which is a well-known open source course management system. From Moodle, the author collected information about student accessing e-learning, where it appeared that some students did not	association, classification, clustering, and outlier detection.	Higher Education	After preprocessing the data, the author applied data mining techniques to discover association, classification, clustering, and outlier detection rules. In each of these four techniques, the author extracted knowledge that describes students' behavior. Author generated association rules and sorted the rules using lift metric then he visualized the rules. Then he generated

			access the system at all. Then, he got information about how much student benefited from resources, such as using e-books, research papers and old exams available on the system. Also, he got the results of students' grades in solving exercizes available in the system.			classification rules using decision tree. Also he clustered the student into group using EMclustering. Finally, by using outlier analysis the author detected all outliers in the data. Each one of this knowledge can be used to improve the performance of student.
M. Abu Tair, A. El-Halees	M. Abu Tair, A. El-Halees (2012). "Mining Educational Data to	Authors in this case study used educational data mining to improve	The data include fifteen years period [1993-2007]. 3360 record of data had	association, classification, clustering and outlier	Higher Education	Results of this case study helped students to increase their performance in

	Improve Students' Performance: A Case Study". International Journal of Information and Communication Technology Research. ISSN 2223-4985, Volume 2 No. 2, February 2012.	graduate students' performance, and overcome the problem of low grades of graduate students. Authors tried to extract useful knowledge from graduate students data collected from the college of Science and Technology – Khanyounis.	been used each record has (Student ID, Student name, Student Address, Student GPA, Gender, DOB, Place of Birth, Enrollment Date, Graduate date, major, phone no., previous information about study history, and semester grade).	detection rules		Higher Education .
E. Osmanbegovi, M. Sulji	E. Osmanbegovi, M. Sulji (2012). "DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE". Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.	to help student and teachers to improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of learning. As learning is an active process, interactivity is a basic elements in this process that affects students' satisfaction and performance.	The data collected from the surveys conducted during the summer semester at the University of Tuzla, the Faculty of Economics, academic year 2010-2011, among first year students and the data taken during the enrollment. After eliminating incomplete data, the sample comprized 257 students who were at the time of researches present at the practice	Different methods and techniques of data mining (Naive Bayes algorithm (NB), Multilayer Perceptron (MLP), decision tree, and neural network methods) were compared in this study during the prediction of	Higher Education	The results of this paper indicated that the Naïve Bayes classifier outperforms in prediction decision tree and neural network methods. It has also been indicated that a gooc classifier model has to be both accurate and comprehensible for professors.

			classes. The model of students' success was created, where success as the output variable is measured with the success in the course "Business Informatics". The success was evaluated with the passing grade at the exam.	students' success. For the purposes of this study WEKA software package was used		

APPENDIX D: SAMPLE OF TRANSCRIPTS DATA

The data investigated represents around 19,800 students in a computer science department at a UK University, For each record, several attributes that represent a student's academic accomplishments at three levels (1st, 2nd and 3rd years) are divided as follows:
Coursework Weighting (Cswk), (year A, year B, year C) (Regno: Registration Number) (Avg: Average)
Module Mark: a student's mark in a certain module
Exam Mark: the mark achieved by a student on the exam-based assessment
Coursework Mark: the mark achieved by a student on the coursework-based assessment.

<>	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Regno	A avg.	B avg.	C avg.	Total Avg	Exam% A	Cswk % A	Both % A	Exam % B	Cswk % B	Both % B	Exam % C	Cswk % C	Both % C
2	Ma9Ghba	61	51.3	34.7058824	49.0019608	0.44444444	0.55555556	0	0.6	0.1	0.3	0.29411765	0.29411765	0.41176471
3	MaRoPPQ	75.6666667	67.8	75.7	73.0555556	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.3	0.3
4	MaCnoZG	71.8181818	71.3	72.7	71.9393939	0.27272727	0.63636364	0.09090909	0.1	0.5	0.4	0.1	0.7	0.2
5	MawGKq9	52.1111111	42.8181818	49	47.976431	0.44444444	0.55555556	0	0.63636364	0.09090909	0.27272727	0.6	0.2	0.2
6	MaREodL	52.375	56.6363636	60.2	56.4037879	0.125	0.375	0.5	0.27272727	0.09090909	0.63636364	0.3	0.4	0.3
7	MaPPGXH	74.2222222	80.1	84	79.4407407	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.2	0.4
8	MaXbX5Y	42.4166667	52.8181818	64.875	53.3699495	0.66666667	0.25	0.08333333	0.45454545	0.18181818	0.36363636	0.25	0.625	0.125
9	MaKjXZI	78.5555556	68.4	72	72.9851852	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.3	0.3
10	Max8aqo	80.4444444	73	71.6	75.0148148	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.2	0.4
11	MafsC8b	61.2727273	63.5454545	69.4444444	64.7542088	0.27272727	0.54545455	0.18181818	0.09090909	0.45454545	0.45454545	0.33333333	0.22222222	0.44444444
12	MaHXv1n	61.7777778	59.3636364	62.4	61.1804714	0.44444444	0.55555556	0	0.54545455	0.18181818	0.27272727	0.4	0.2	0.4
13	MamgS20	73.2222222	58.9090909	69.9	67.343771	0.44444444	0.55555556	0	0.54545455	0.18181818	0.27272727	0.3	0.4	0.3
14	Ma5IlG6	64.4545455	71.8	74.6	70.2848485	0.27272727	0.63636364	0.09090909	0.1	0.5	0.4	0.1	0.6	0.3
15	Ma7rjTO	72.4444444	73.8	77.8	74.6814815	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.3	0.3
16	MaJcGbu	55.0833333	55.6	69.3333333	60.0055556	0.33333333	0.58333333	0.08333333	0.1	0.6	0.3	0.11111111	0.55555556	0.33333333
17	Ma5RIGh	68.1111111	65.0909091	6.4	46.5340067	0.44444444	0.55555556	0	0.54545455	0.18181818	0.27272727	0.4	0.4	0.2
18	MaO51mH	82.2222222	70.3	77.2	76.5740741	0.44444444	0.55555556	0	0.6	0.1	0.3	0.3	0.4	0.3
19	MaFU6od	41.5555556	52.2727273	60.6666667	51.4983165	0.77777778	0.11111111	0.11111111	0.54545455	0.09090909	0.36363636	0.22222222	0.55555556	0.22222222
20	Ma186Wp	69.7272727	62.5454545	70.25	67.5075758	0.72727273	0.18181818	0.09090909	0.54545455	0.09090909	0.36363636	0.5	0.375	0.125
21	MaCHYjQ	58.6363636	59.9090909	68.6666667	62.4040404	0.72727273	0.18181818	0.09090909	0.54545455	0.18181818	0.27272727	0.33333333	0.22222222	0.44444444
22	Ma3nKXQ	72.2222222	50.6363636	61.6	61.4861953	0.44444444	0.55555556	0	0.54545455	0.18181818	0.27272727	0.4	0.2	0.4
23	Ma8vb4D	72.4444444	68.6	74.4	71.8148148	0.44444444	0.55555556	0	0.6	0.1	0.3	0.4	0.2	0.4
24	Mam58iL	67.0909091	61.8181818	64	64.3030303	0.72727273	0.18181818	0.09090909	0.54545455	0.09090909	0.36363636	0.625	0.25	0.125
25	Ma9IJyy	66.0909091	57.1818182	57.6666667	60.3131313	0.72727273	0.18181818	0.09090909	0.45454545	0.18181818	0.36363636	0.55555556	0.22222222	0.22222222
26	MaK08da	72.4545455	68.2727273	70.25	70.3257576	0.72727273	0.18181818	0.09090909	0.54545455	0.09090909	0.36363636	0.5	0.375	0.125
27	Mac5EfQ	61	63.3636364	63	62.4545455	0.72727273	0.18181818	0.09090909	0.45454545	0.18181818	0.36363636	0.25	0.625	0.125
28	MavaOYF	78.25	72.8181818	69.1	73.3893939	0.125	0.375	0.5	0.27272727	0.09090909	0.63636364	0.4	0.3	0.3
29	Ma3xFnW	52.5	52.8181818	68.875	58.0643939	0.75	0.16666667	0.08333333	0.54545455	0.09090909	0.36363636	0.25	0.625	0.125
30	MavEWvy	72.2222222	64	69.7	68.6407407	0.44444444	0.55555556	0	0.5	0.2	0.3	0.4	0.3	0.3
31	Ma3kBfN	43.7272727	40	60.3	48.0090909	0.54545455	0.45454545	0	0.58333333	0.16666667	0.25	0.2	0.6	0.2
32	Ma3RLry	63.6666667	62.1	70	65.2555556	0.44444444	0.55555556	0	0.5	0.2	0.3	0.2	0.4	0.4
33	Maj1i1v	56	56.1818182	61.2	57.7939394	0.44444444	0.55555556	0	0.63636364	0.09090909	0.27272727	0.4	0.2	0.4
34	MaM7juj	49.0909091	40.2142857	20.7	36.6683983	0.36363636	0.63636364	0	0.5	0.21428571	0.28571429	0.6	0.1	0.3
35	Ma58sLi	83.6666667	68.9090909	68.7	73.7585859	0.44444444	0.55555556	0	0.54545455	0.18181818	0.27272727	0.3	0.4	0.3
36	MaqKaGs	59.6363636	60.4545455	65.1	61.730303	0.27272727	0.54545455	0.18181818	0.09090909	0.36363636	0.54545455	0.1	0.7	0.2
37	MajjRgQ	63.8888889	50.3	63.4	59.1962963	0.44444444	0.55555556	0	0.5	0.2	0.3	0.3	0.3	0.4
38	Maw5RAq	66.5555556	72.1	68.3	68.9851852	0.44444444	0.55555556	0	0.6	0.1	0.3	0.3	0.4	0.3
39	MagYcLV	42.2142857	49.9090909	58.375	50.1661255	0.71428571	0.21428571	0.07142857	0.45454545	0.18181818	0.36363636	0.375	0.5	0.125
40	MazIRYh	55	56.5	63.8	58.4333333	0.5	0.5	0	0.6	0.1	0.3	0.2	0.6	0.2
41	MarQHrH	52.4444444	53.6666667	52	52.7037037	0.11111111	0.33333333	0.55555556	0.25	0.08333333	0.66666667	0.4	0.2	0.4
42	MaOKSJF	68.4545455	67.6363636	70.4444444	68.8451178	0.72727273	0.18181818	0.09090909	0.45454545	0.27272727	0.27272727	0.33333333	0.44444444	0.22222222
43	MazX4oR	44.8	65.7272727	69.5	60.0090909	0.2	0.4	0.4	0.27272727	0.09090909	0.63636364	0.4	0.2	0.4
44	MaG9RFd	70.5454545	71.2727273	67.6	69.8060606	0.27272727	0.54545455	0.18181818	0.09090909	0.36363636	0.54545455	0.3	0.5	0.2

APPENDIX E: SAMPLE OF ATTENDANCE DATA

The student attendance data represent data from 59 modules for 19,800 students that were gathered from Computer Science undergraduates of a UK university during the years 2010-2015. The main attributes of this table consist the attendance status, the attendance reason, module code and the attendance date. The chosen columns were the attendance status and attendance date to be processed for this thesis.

	A	B	C	D	E
1	Regno	attended	attended_reas	module_code	attended_date
2	MaG4Ka9	y	1	10DSA004	10/5/10
3	MaMOMAs	y	1	10DSA004	10/5/10
4	MaO1KMu	y	1	10DSA004	10/5/10
5	MaddTu1	y	1	10DSA004	10/5/10
6	MadYNGS	y	1	10DSA004	10/5/10
7	Maqwo1T	y	1	10DSA004	10/5/10
8	Majyxld	y	1	10DSA004	10/5/10
9	Man64cW	y	1	10DSA004	10/5/10
10	Mavvhe6	y	1	10DSA004	10/5/10
11	MaPeb4a	y	1	10DSA004	10/5/10
12	Mak8wMM	y	1	10DSA004	10/5/10
13	MaRhNOH	y	1	10DSA004	10/5/10
14	MaD51SB	y	1	10DSA004	10/5/10
15	MaIn0TF	y	1	10DSA004	10/5/10
16	Ma9Tdz4	y	1	10DSA004	10/5/10
17	Ma9KIEJ	y	1	10DSA004	10/5/10
18	MawfEw6	y	1	10DSA004	10/5/10
19	MacpTRz	y	1	10DSA004	10/5/10
20	Ma9dp56	y	1	10DSA004	10/5/10
21	MaR3AzQ	y	1	10DSA004	10/5/10
22	MaC5URf	y	1	10DSA004	10/5/10
23	Ma5iRcI	y	1	10DSA004	10/5/10
24	MaGIm2C	y	1	10DSA004	10/5/10
25	MaIHGb4	y	1	10DSA004	10/5/10
26	MaJNGSs	y	1	10DSA004	10/5/10
27	MaIJHz0	y	1	10DSA004	10/5/10
28	MatYzyL	y	1	10DSA004	10/5/10
29	MaEOpE4	y	1	10DSA004	10/5/10
30	MaKOPGU	y	1	10DSA004	10/5/10
31	MajT66f	y	1	10DSA004	10/5/10
32	MaO2Fsk	y	1	10DSA004	10/5/10
33	MaQw6bz	y	1	10DSA004	10/5/10
34	MaAWiut	y	1	10DSA004	10/5/10
35	MaIRDKA	y	1	10DSA004	10/5/10
36	Ma998vk	y	1	10DSA004	10/5/10
37	May7Soz	y	1	10DSA004	10/5/10
38	Ma6dbwv	y	1	10DSA004	10/5/10
39	MafjE72	y	1	10DSA004	10/5/10
40	MaXBAmW	y	1	10DSA004	10/5/10

APPENDIX F: SAMPLE OF PREPARED DATA

Final modified data were used in this thesis is the number of records for computer science students in the undergraduate level which is 788 records. The columns of this table consist the module mark of level A, the module mark of level B, the module mark of level C, the total average of all levels, the refined total average, SAC average and CAR average.

	Regno	Module Mark A	Module Mark B	Module Mark C	Total Average	Refined Total Average	AVERAGE SAC	Average CAR
1	Ma0AGRn	47.83	51.18	63.44	57.31	56.42	0.49	0.30
2	Ma0cR6k	63.78	41.83	54.20	48.02	47.52	0.34	0.25
3	Ma0DXR7	55.64	53.10	67.90	60.50	58.47	0.59	0.50
4	Ma0G6q4	70.78	69.40	80.20	74.80	74.30	0.85	0.25
5	Ma0GHl3	53.89	60.80	61.80	61.30	60.40	0.66	0.30
6	Ma0Mlsl	79.56	74.40	79.10	76.75	76.25	0.73	0.25
7	Ma0pLFm	63.89	53.18	69.10	61.14	60.24	0.72	0.30
8	Ma0YyXf	52.38	61.09	66.90	64.00	63.49	0.78	0.25
9	Ma186Wp	69.73	62.55	70.25	66.40	65.50	0.72	0.30
10	Ma19hPR	41.29	44.33	42.50	43.42	40.65	0.53	0.55
11	Ma1axiU	70.89	64.10	73.30	68.70	68.20	0.66	0.25
12	Ma1dHYz	53.91	53.00	65.33	59.17	54.59	0.56	0.65
13	Ma1o2w2	56.64	58.82	65.11	61.96	58.35	0.53	0.60
14	Ma26J8p	71.56	64.00	64.20	64.10	63.20	0.75	0.30
15	Ma2a8Zc	56.45	58.82	62.56	60.69	59.79	0.71	0.30
16	Ma2AQQG	60.50	53.00	66.40	59.70	59.20	0.53	0.25
17	Ma2GZRW	47.17	51.09	62.33	56.71	56.49	0.52	0.20
18	Ma2Jybk	68.27	65.18	66.88	66.03	64.62	0.58	0.40
19	Ma2WdU8	90.33	80.60	82.20	81.40	80.90	0.81	0.25
20	Ma32vK5	69.82	68.36	73.56	70.96	70.46	0.74	0.25
21	Ma32xQt	44.08	55.36	63.89	59.63	58.22	0.62	0.40
22	Ma3Cd19	68.27	68.45	64.44	66.45	65.55	0.51	0.30
23	Ma3h8FR	53.36	49.75	50.33	50.04	46.43	0.72	0.60
24	Ma3iGkd	69.00	54.00	63.90	58.95	58.05	0.56	0.30
25	Ma3kBfN	43.73	43.64	60.30	51.97	50.56	0.63	0.40
26	Ma3nKXQ	72.22	50.64	61.60	56.12	55.22	0.52	0.30
27	Ma3RLry	63.67	62.10	70.00	66.05	64.65	0.74	0.40
28	Ma3waER	65.50	72.27	77.50	74.89	74.38	0.78	0.25
29	Ma3xFnW	52.50	52.82	68.88	60.85	59.44	0.76	0.40
30	Ma441yP	64.55	53.70	2.04	27.87	25.11	0.23	0.55