Predictive modelling of student academic performance – the case of higher education in Middle East

A thesis submitted in partial fulfilment of the

requirements of the University of East London

for the degree of Professional Doctorate in Data Science

School of Architecture, Computing & Engineering

University of East London

By **Wesam Al Madhoun**

Professor: Allan Brimicombe

London, United Kingdom

2020

Acknowledgment

First, and most, I would like to express my most profound appreciation to my primary supervisor and the Director of Studies, Professor Allan Brimicombe for his continuous support, help, advice, and encouragement. The opportunity I had to learn from him is adding value to my knowledge. I would also like to thank Dr. Yang Li, my second supervisor, and the one who was a great help from the first day I applied to join the University. Special thanks to my parents and my family in general for their support, encouragement, and understanding. Also, I would like to thank Qatar University for providing the data needed for this project.

Thanks to everyone who had helped to accomplish this dissertation which I would not have finished without their support

Dedication

I dedicate this project to my family and especially to my father, who is always encouraging me to achieve a level of excellence.

Abstract

One of the main issues in higher education is student retention. Predicting students' performance is an important task for higher education institutions in reducing students' dropout rate and increasing students' success. Educational Data mining is an emerging field that focuses on dealing with data related to educational settings. It includes reading the data, extracting the information and acquiring hidden knowledge. This research used data from one of the Gulf Cooperation Council (GCC) universities, as a case study of Higher Education in the Middle East. The concerned University has an enrolment of about 20,000 students of many different nationalities. The primary goal of this research is to investigate the ability of building predictive models to predict students' academic performance and identify the main factors that influence their performance and grade point average. The development of a generalized model (a model that could be applied on any institution that adopt the same grading system either on the Foundation level (that use binary response variable (Pass/ Fail) or count response variable which is the Grade Average Point for students enrol in the undergraduate academic programs) to identify students in jeopardy of dismissal will help to reduce the dropout rate by early identification of needed academic advising, and ultimately improve students' success.

This research showed that data science algorithms could play a significant role in predicting students' Grade Point Average by adopting different regression algorithms. Different algorithms were carried out to investigate the ability of building predictive models to predict students' Grade Point Average after either 2, 4 or 6 terms. These methods are Linear/ Logistic Regression, Regression Trees and Random Forest. These predictive models are used to predict specific students' Grade Point Average based on other values in the dataset. In this type of

model, explicit instruction is given about what the model needs to learn. An optimization function (the model) is formed to find the target output based on specific input values.

This research opens the door for future comprehensive studies that apply a data science approach to higher-education systems and identifying the main factors that influence student performance.

List of Figures

List of Tables

Table of Contents

Glossary

AGPA	Accumulative Grade Point Average
ANN	Artificial Neural Network
CART	Classification And Regression Trees
CHID	CHI- Squared Automatic Interaction Detector
DM	Data Mining
EDM	Educational Data Mining
FN	Foundation
GCC	Gulf Cooperation Council
GPA	Grade Point Average
IT	Information Technology
ITS	Information Tutoring System
KDD	Knowledge Discovery in Database
KEEL	Knowledge Extraction Evolutionary Learning
KN N	K Nearest Neighbour
LM	Learning Analytics
MIL	Multiple Instance Learning
ML	Machine Learning
MOOC	Massive Open Online Course
QUEST	Quick Unbiased and Efficient Statistical Tree
Rattle	R Analytical Tool to Learn Easily
RMSE	Root Mean Square Error
SIS	Student Information System
UG	Undergraduate
WEKA	Waikato Environment for Knowledge Analysis

Chapter 1 Introduction

1.1 Research motivation

One of the significant challenges facing academic institutions is student retention to degree completion. The ability to predict students' performance may increase student retention by identifying academic advising needs at an early stage. Because of the massive amount of student data generated at higher education settings, there is a need for a mechanism to read and analyze the historical data to predict the student's performance. According to Trcka (2010), traditional data mining has been widely used to extract hidden patterns from the massive volume of datasets.

One of the most efficient techniques to extract this hidden knowledge is data mining (DM) and machine learning (ML) techniques. According to Kamath and Karat (2016), data mining technology is widely used in educational fields to predict the main factors that play a significant role in students' performance. That could be done by using the collected data about students' records and test scores, as well as non- academic information, to build models to predict students' academic performance.

One of the major problems facing higher education administrators today is student attrition. Many students who enter higher education leave without earning a degree. Research into the causes of attrition as it affects student retention and success is institution specific (Upcraft, Gardner & Barefoot, 2005). Based on the number of variables and attributes identified through these investigations, researchers have developed many formulas for student success. Nevertheless, the applicability of these research results needs further consideration when

examined within the context and the culture of institutions in other parts of the world. Research on student retention in the Gulf-Arab countries is scarce, making it even more important that each institution establish its own research efforts into the causes of attrition.

Cultures in the Gulf-Arab region are rather collective in nature (Hofstede & Hofstede, 2005). Living with their parents after the age of 18 is a cultural norm for students in the region. Therefore, most students in the Gulf-Arab region who enrol in college live with their parents. Another substantial number of students is married and has children and continues to go to college. This results in a large number of commuter students in higher education institutes in the Gulf-Arab region.

The impetus for such a concern is the large-scale changes that have been experienced within the field of higher education in the past few decades. The changes in the field of higher education have led to an enhanced cultural diversity. People of various educational, as well as cultural, backgrounds have been integrated into higher education. In the past, when there was greater homogeneity in terms of the student profiles, academic completion, as well as success, was seen more as the student's responsibility. Presently though, given the diversity of student cohorts, it has become imperative for education institutions also to take responsibility for student's performance and retention. More generally, the better the academic programs provided by a given university, the more the likelihood that students will be retained, resulting in enhanced revenue by the university.

According to Oussena, (2008), it is most likely for students to stay on a given course if there are close links between the academic characteristics of a given university and their personal, educational goals and objectives. Dropouts and poor performance have been majorly rampant

particularly because students find it hard to combine their chosen subjects and the academic characteristics of a given educational institution. Most of the problems leading to poor performance and poor retention for students are related to poor grounding for higher education and mismatches between the institutions and the student. This usually leads to inconsistency between the values of the students and those of the institution. It is because of the above factors that data mining is essential to help educational institutions come up with ways to improve student performance and retention.

Delavari *et al.* (2008) touched on the knowledge gap and a lack of significant information on counselling, student registration, and the grading process. Delavari *et al.* (2008) presented a case study of adopting data mining techniques in higher-learning institutions. Although this work showed a case study of utilizing data mining techniques in a particular course, Computer Programming II, and its prerequisite Computer Programming I, it is impressive that the authors proposed comprehensive new guidelines called Data Mining in a Higher Education System (DM_EDU). These models are different from the traditional DM-CRISP in the sense that its unique elements provide a more effective and enhanced roadmap through which data mining should be done within a higher educational system. The main components proposed in these new guidelines were evaluation, planning, registration, consulting, marketing, and performance. For each element, the authors identified the detailed sub-processes and applied data mining techniques. They designed descriptive and predictive models to determine the relationships between factors that affected student's performance in the specific course mentioned above. Oussena (2008) notes that the decision regarding whether a student drops out or continues with education is quite strongly linked to the degree to which they are academically and socially integrated (Schendel and McCowan, 2016). The study indicates a positive correlation between social and academic integration and student performance and

retention. Poor student retention and associated performance are in most cases due to unclear career aims, uncertainty concerning goals, adjustment or transition challenges, and lack of academic motivation as well as unrealistic or limited expectations.

One of the critical areas is building a model that helps in predicting students' knowledge, helping to improving students' performance (Romero et al., 2007). Another important area is how the pedagogical and learning management systems could be used in improving learners' skills and improving their performance in future both currently an in the future.

The institution selected for study is one of the gulf national universities, and therefore the first choice for students from this country. About 20,000 students are enrolled in various undergraduate academic programs and colleges. It hosts ten colleges -- College of Arts and Sciences; College of Business and Economics; College of Education; College of Engineering; College of Health Sciences; College of Law; College of Medicine; College of Pharmacy; College of Sharia and Islamic Studies and College of Dental Medicine. It offers a wide range of academic programs -- 45 Bachelors, 27 Masters, eight Ph.D. programs, four Diplomas, and a Doctor of Pharmacy. However, data indicates that student retention falls off after students have completed more than 50% of the degree requirements. In its Strategic Plan (2018-2022), it seeks to" Improve students' academic success throughout the whole student lifecycle from pre- university stage to beyond the graduation" (p. 21). Results from this study may help to achieve this objective. The goal of this study is to build a predictive model(s) by using data mining and machine learning techniques, to predict students' Grade Point Average (GPA), based on historical academic / non-academic data available in the university's Student Information System (SIS), and the Information Technology (IT) department.

1.2 Background

According to Nisbet, Elder & Miner (2009, p. 17) data mining was defined as "the use of machine learning algorithms to find patterns of relationship between data elements in large, noisy, and messy dataset, which can lead to actions to increase benefit in some forms (diagnosis, profit, detection, etc)". Educational Data Mining (EDM) is a computer-based information system used and devoted to scanning massive amount of educational data, generating information and discovering institutional knowledge (Anjewierden, 2011). As indicated by Ismail & Abdulla (2015), some of the notable emerging Data Mining (DM) applications are predicting students' performance, students' modelling, and recommendations for students and planning and scheduling. The authors pointed out that in general, the objective of predicting students' performance is to predict unknown variables that might affect students' during their academic journey.

Asif et al. (2014) listed some everyday tasks related to educational data mining to predict performance on different levels. These were highlighted by Romero & Ventura (2010) and include predicting students' performance at the tutoring system level, on the course level or on the degree level. It could also be used to predict student's grades, or at tutoring system to predict if the student will get the next question or training exercise, or to predict if the student will pass a course based on his/ her activities or combination on the course forum. These are some common tasks on different levels to predict students' performance.

1.3. Need for Study

Public and private universities have operated in the Middle East for decades, however the number of institutions is insufficient to meet the population demand. Further, exceptional

Middle East universities fail to adequately equip students with the skills needed for the local or global job market; and the scientific higher education and research have been overlooked. In recent years, Arab governments and educational elites have constantly the importance of higher education and carried out various initiatives that can bring out substantial changes. States belonging to the Gulf Cooperation Council (GCC) including Kuwait, Saudi Arabia, Bahrain, Qatar, the United Arab Emirates, and Oman, have invested billions to the development of new institutions in the past decade. (QS Asia News Network, Sept. 14, 2018).

An essential aspect to increasing the quality of higher education in the Middle East is improving student outcomes, which is dependent upon identifying struggling students by providing needed support in a timely manner. Students of any undergraduate program can be placed under academic probation based on the cumulative Grade Point Average (GPA) if their GPA is below 2.00 out of 4.00. If the student placed on academic probation for two consecutive terms (without counting summer term) has failed to raise the cumulative GPA to 2.00 at the end of the following term, then the student placed under final probation. As per to the university's catalog (2018/ 2019, P. 95-96), once the student is placed on final academic probation and failed to raise the GPA to 2.00 or failed to finish all degree requirement within eight years from enrolment, then he will get dismissed from the university.

The implication of failure raising the GPA, and dismissed are affecting the university's image as well as the student's retention as student retention is an important topic not just for parents, but also for higher education administrators, instructors and every stakeholder involved in the educational process. The more the students retained, the more like hood the country meet their future needs from graduates in all disciplines who help in the country's future vision.

As per of Qatar National vision 2030, p.11, one of the country's vision pillars is "Development of all its people to enable them to sustain a prosperous society" . In order to achieve this, there is a real need for well-equipped generations to serve and meet future needs for all pillars (Human development, Social development, Economic and Environmental development) listed in the country's vision.

The main aspiration of this research is to build a predictive generalized model (a model that could be applied on any institution that adopt the same grading system either on the Foundation level (that use binary response variable (Pass/ Fail) or count response variable which is the Grade Average Point for students enrol in the undergraduate academic programs) to predict student's GPA. Thich will help Middle East Institutes in general, and the concerned University in particular, to identify at-risk students at the early stage before dismissal from the university.

1.4. Research Site

The concerned university is the country's first higher education institution and is located in the capital city Doha. It hosts ten college. It offers the widest range of academic programs -- 45 Bachelors, 27 Masters, eight Ph.D. programs, four Diplomas, and a Doctor of Pharmacy (PharmD)-- in the country tailoring them to meet the needs of society. There are approximately 20,000 students enrolled. This study focused on Undergraduate students.

All students are expected to possess minimum basic skills to be eligible for enrolment in their desired academic programs. In order to be considered for Undergraduate admission to the university, all applicants applying to the following colleges must demonstrate proficiency in English and Mathematics by satisfying the following minimum competency requirements as set by the University or pass the Foundation Program. The University's Foundation Program is

an academic entry program designed to bridge any potential gaps between the student's minimum academic skills upon graduating from secondary school and the academic level needed to be successful at the University level.

1. 5. Data Mining and Machine Learning

Since both concepts deal with data, and it looks like that both are the same. Although as indicated by Lantz (2015), one of the potential points of distinction is that Machine Learning is focusing on teaching the computer to use data to solve a problem, while data mining is showing the network to extract the knowledge from hidden information and to identify patterns that human use to solve a problem.

Machine learning techniques are divided into two types. These are:

- **Supervised Learning**: in this type **predictive** models are used to predict specific target based on other values in the dataset. In this type of model, explicit instruction is given about what the model needs to learn. An optimization function (the model) is formed to find the target output based on specific input values. One of the often used in the classification.
- **Unsupervised Learning**: descriptive models are used. According to Lantz (2015), no single feature or target value to learn, while it is all about pattern discovery which used to identify the associations within the dataset values.

  Since the main purpose of this work is to predict students' GPA, the focus will be on the first type of Machine Learning which is supervised learning by using different algorithms to build predictive models to predict the student's future GPA based on certain historical values. Four supervised learning algorithms used and compared to find out the best method to predict the student's GPA. These are:

  - Linear/ Logistic Regression

- Regression Trees
- Random Forest

The following section contains a brief explanation of each technique.

1.6. Supervised Learning algorithms

- **Linear Regression**

Linear regression is used for not just classification, but also for predicting a numerical value and estimating the relationship between these numerical values. This relationship is between the target (output) variable, and more than one independent variable or predictors. In case more than one independent variable used, it is called multiple linear regression which is the case in the present work.

Regression can also be used for other types of dependent variables. For instance, according to Lantz (2015), logistic regression is used to model binary categorical outcomes. O.D.& P.A (2017) presented multiple regression models to predict the student's academic performance (SAP) using data of the Department of Computer Science in Nigeria. In this type of regression, the regression equation models data by using similar slope- intercept format as $y = a + bx$. In this case, the machine's job is to find the best values of a and b, so the line is best able to fit and relate to the proved x values to the value of the target variable.

- **Regression Trees**

  Decision trees are one of the supervised learning algorithms that used widely for data exploration purposes. According to Yadav et al. (2012), a decision tree is a flow- chart like a tree structure. Rectangles denote each interval node, and ovals indicate each leaf nodes. There are different two main types of decision trees. These are:

  - Classification tree: used if the response (target) variable is categorical, and

- Regression tree: used if the target variable is a real numerical number as it is in our case in the present work (Student's GPA).

Koksiantis & S.B (2012) presented a decision support system for predicting student's grades. The author pointed out that although the model trees are small and more accurate than regression trees, as indicated by Wang & Witten (1997), regression trees are comprehensible.

Ibrahim & Rusli (2007) compared the machine learning algorithms (linear regression, Artificial Neural Network (ANN), and decision tree) to predict the performance of 206 undergraduate students. The result showed that the ANN model outperforms than the other two algorithms. This study was restricted on relatively small size dataset.

- **Random Forest**

  Random forest as one of the supervised machine learning algorithms. It is considered as a collection of decision trees. It is an ensemble learning technique for both classification and regression tasks, which make it widely used. Furthermore, one of the main advantages of random forest algorithms is that it is easy to measure the importance of the model's features or variables by computing the score automatically for each node of the tree.

  Compared with a decision tree, in the decision tree, a set of rules are generated based on the features that describe the tree, while in random forest algorithm, observations are selected randomly to build several decision trees, and then takes the average of the results.

1.7. Outline of the thesis

The present thesis is organized into eight chapters described as follows:

Chapter 1 introduces the problem for study, describes the rationale and purpose of this study, background information and specific site information and outlines the study for the reader.

Chapter 2 reviews the professional literature related to the variables under study such as learning analytics, data mining algorithms, and predictive factors. The implications of the literature review for the current study lead to a delineation of the research objectives and research questions. The significance of the proposed study is discussed

Chapter 3 presents the research methodology and design, including ethical considerations. Describes sources of data, data collection and analysis procedures. used to achieve the main thesis objectives by answering the research questions that were listed in the previous section.

Chapter 4 describes machine learning, and data mining. It introduces the data science algorithms and models built to predict student's performance based on GPA.

Chapter 5 presents the implementation of machine learning algorithms for students enrolled in the foundation program. The sample is described, and the results of factor analysis and logistic regression are presented. More data science algorithms are discussed.

Chapter 6 discusses the implementation of machine learning algorithms on undergraduate students. Preparations on the dataset are discussed, as are variable transformations. Varying groups of students are compared. An evaluation of the results of machine learning algorithms is presented.

Chapter 7 presents testing the correlation between the Foundation program scores and students' GPA after declaring their majors.

Chapter 8 presents a discussion of the findings, discussion, and contribution of this research, limitation and areas for future research.

Chapter 2 Review of Literature

2.1. Introduction

This chapter first presents a critical review of the professional literature involving predicting students' performance as well as the Machine Learning (ML) techniques used to predict students' performance at different stages or levels. This provides the basis to formulate research questions.

2.2 Learning Analytics

Predicting student's performance is an essential task for higher-education institutions in reducing student's dropout rate. This section addresses different ways of using Data Mining and Machine Learning techniques to predict students' academic performance, with a view to improving the student learning environment. Main themes have been identified in order to group concepts together. The following structure illustrates the review steps as follows:

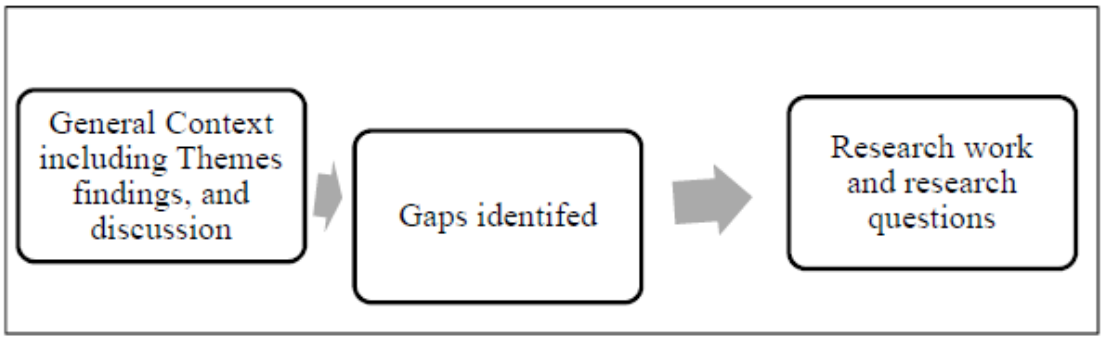


Figure 2. 1 Review steps

The principal purpose of this study is to identify the main factors that affect students' performance and examine which algorithms do better in predicting students' performance. This

review will focus on related research work on two themes related to the following objectives. These are:

- Investigating the ability of building a predictive model to predict students' performance
- Factors affect students' performance

2.3 Data Mining in Higher Education

Olson, (2007) defines data mining as the process which entails the uncovering of patterns and discovery of anomalies as well as relationships associated with large datasets which can be applied in the predictions relating to future trends. The fundamental purpose of data mining is to extract valuable information from the available data (Stephenson, 2016).

Huebner (2013) conducted a survey of EDM that explored different data mining techniques, such as clustering, classification, and association rule mining, but data mining has not been given as much attention in education as other areas of research. Although the survey covered the vital work done in the field of EDM, the limitation is that most of the research highlighted in this paper was focused on case studies that fit with specific organizations or institutions and are difficult to generalize.

Data mining techniques have been applied on a different and wide range of fields and disciplines, whereas Data Mining in Education is an emerging interdisciplinary field that is used to discover hidden knowledge from educational data. Educational data mining (EDM) has been made more accessible and feasible with the introduction of public education data repositories in 2008. Data mining has been used in various aspects of higher education.

Romero and Ventura (2007) have raised the importance of adopting data mining techniques in the educational field in e-learning and web mining. They presented a new model for using data mining technology in higher-education systems, text mining, etc., and showed that data mining was a new and promising area of research. One of the strengths of this research work is that it made an important point that not many other authors raised, which is that most data mining techniques are not easy to use by non-experts, and so there is a real need for simple procedures that serve teachers' needs. The same point was highlighted by El Atia and Hammad (2012) and Dogan and Camur (2008).

There are different applications of educational data mining. Vahdat *et al.* (2015) and Papamitsiou and Economides (2014) discussed the issue of integrating learning analytics and data mining in the educational setting to improve the learning process and enhance educational systems. However, Vahdat et al.'s (2015) work is much more comprehensive and presents more applications for adopting learning analytics and data mining in the educational field.

Beikzadeh *et al.* (2004) analyzed a model that represents the data mining process in education. The decentralized model includes the main steps of educational systems: planning, evaluation, and counselling. A sub-process was presented for each part, and the built model can be used to improve the educational system in higher-education institutions. Although this work is comprehensive in terms of providing a variety of applications for data mining techniques in the educational field, some factors affected the reliability of the designed model. For example, some missing values and attributes were ignored.

Delavari *et al.* (2004) created a new model for using data mining technology in higher-education systems. The model proposed by Delavari *et al.* (2004) is referred to as Data Mining for Higher Education (DM_EDU). The model enhanced the description of how Data Mining can be applied in systems of higher education to increase the productivity of the traditional

processes. It provides a guideline for the operation of decision making. The research work focused on improving student performance, course assessment, and significant selection.
Dogan & Camur (2008) introduced and demonstrated the use of data mining techniques to predict student performance and to extract their mistakes in an intelligent tutoring system (ITS).

Tair and El Halees (2012) focused on graduate records in a case study that adopted data mining techniques to discover knowledge from educational data and improve graduate students' performance. The study covered 3341 records from 1993 to 2007. Different data mining techniques were deployed, and it was pointed out that the predicted model provided significant results that could be used to improve graduate students' performance.

El Atia and Hammad (2012) addressed the use of data mining techniques in improving the learning system in Canadian institutions. The authors identified a lack of a systematic approach for collecting, storing, and analyzing the data. Furthermore, most research work done in the field of educational data mining was mainly done by computer/engineering researchers. So, there is a need for more collaboration with experts in education to bridge that gap.

Khan and Choi (2014) presented an interesting educational data mining application for scholarship prediction. They designed a calculator to tabulate a student's chance of winning a scholarship based on the entered information. In my opinion, although the title of this paper does not relate to the proposed work, the approach of this paper was easy to understand, and it was accurate. The authors highlighted the importance of using the ID3 decision trees compared with C4.5 since ID3 has more rules and depth, providing much more accuracy for the built model.

The need for data mining techniques discussed above is therefore critical given the role they play in educational research concerning students' performance and retention. These techniques are helpful in the prediction of the learner's behaviours and thereby facilitating the improvement of student models. Student modelling entails the characterization of students in terms of their knowledge, meta-cognition, and attitudes as well as motivation. The employed data mining techniques are further used in the discovery and improvement of knowledge domain structures with educational institutions. They also provide a framework for the study of some of the most operational pedagogical support for the learning of students which can be attained through learning systems. The rationale for these data mining techniques is also based on their capacity to establish empirical evidence necessary for the support and articulation of pedagogical theories, educational phenomena, and frameworks essential in determining the core learning components needed for the design of the better system of learning. Some of these included the Pittsburgh Science of Learning Centre's (PSLC) Data Shop as well as the National Center for Education Statistics (NCES).

2.4 Theme 1: Related work on Data Mining algorithms

As one of the main objectives of this work is to identify the best algorithms that could be used in predicting students' performance, the following discussion will address the related work on data mining algorithms used for the same purpose.

Vandamme *et al.* (2007) produced a model for predicting student performance and for classifying students at an early stage into three categories: low-risk students who have a high probability of succeeding; medium-risk students who might succeed; and high-risk students who have a high likelihood of failing. The experiment involved 533 students enrolled in three

universities (Jiawei & Micheline, 2006). The difficulty was in classifying the first-year students into the three categories mentioned above before their first university examinations. Although the prediction obtained was not so good, the linear discriminate analysis performed better compared with other algorithms. The model needs to be enhanced by applying it to a more significant number of students to test their performance.

Hung and Crooks (2009) presented a study that addressed the impact of the learning management system on improving student performance. Three data mining techniques (clustering analysis, association rule, and decision trees) were applied to examine and extract knowledge and patterns from peer-moderated and teacher-moderated groups. The results appeared more coherent and accurate than those attained by Tsai et al. (2011). The results showed that data mining techniques are useful tools in identifying and predicting students' behaviour, and there is a need for verifying these models and to integrate them in the learning management systems so that they can send reminders to teachers and students. Although this is a complete work, there is a limitation as all data mining algorithms were applied to a small number of students (98 first-year college students). One unique point this work considered was that the built model found the demographic information and high school scores on the SA1T, ACT, etc. The research data sets were tested on four data sets extracted from George Mason University (GMU), the University of Minnesota (UMN) Learning Management System, and Stafford University MOOC data. The outcomes showed that the built model could be used as a system-based personalized analytics system to forecast student performance correctly, and the error rate is less than that of traditional methods. Furthermore, there is a need for refining these models, so they can be used in degree planning to solve the retention problem.

Kovacic (2010) studied the impact of using the enrolment data to predict students' success in the Open Polytechnic of New Zealand. The data set of 450 students' records stored in the student management system from 2006 to 2009 were examined in this work. A CART classification tree was applied. Ethnicity was shown to be one of the main factors affecting student performance. Furthermore, the authors raised the issue of the importance of the impact of the semester when the singular course was offered on student performance.

This work highlighted the influence of ethnicity on performance; however, the accuracy of the classification method was 60.5%, which is low. The study was carried out on 450 students in a specific course, Information Systems. Finally, the focus was on non-academic attributes, and there is no evidence as to how other academic backgrounds could influence the model.

Veeramuthu *et al.* (2014) studied how different factors affect students' learning and performance using clustering as one of the data mining techniques to predict students' results. Similarly, Tsai *et al.* (2011) presented a case study that applied clustering techniques at the National University in Taiwan to cluster and predict the undergraduate students who might fail the computer proficiency test as part of the graduation requirements at the university. The model in this paper helps the university develop an early-warning system to identify these categories. The K-means algorithm was applied along with unsupervised neural networks. Finally, the C5.0 decision-tree algorithm was used, and several decision rules were extracted to predict student performance on other universities' tests, such as English language tests, that are required at the same institutions. The K-means algorithm appeared to be the most effective in terms of the results obtained compared to neural networks and C5.0 decision-tree algorithm.

Natek and Zwilling (2014) presented a new study that used three decision tree techniques on small data sets to predict students' success rate. Romero *et al.* (2008) introduced the theoretical and practical approach of applying data mining algorithms (statistics, visualization, classification, and association rules) for Moodle's data as a learning management system. Two free and open sources were used: WEKA and KEEL. For this purpose, and two data sets were mined. Female students performed better than male students but were restricted to certain majors (Hussain and Hazarika, 2014). A similar approach was produced by Romero *et al.* (2008) but focused on comparing different classification algorithms. A specific mining tool was integrated into Moodle so instructors can use these tools. A data set of 438 students' records related to seven Moodle courses offered by Cordoba University was used.

Asif *et al.* (2014) used the same data mining algorithms used by Vandamme *et al.* (2007)—decision trees and neural networks—but added Naive Bayes. The focus was on decision trees and rule induction algorithms since these two methods are easy to interpret and both can be used to discover the courses in the first and second year. Without using the demographic information, the classifier can predict graduation performance with high accuracy compared to other studies.

The same techniques were applied by Abu-Oda and Abu-Oda (2015) who did not develop new methods of predicting and classifying student performance but used classifications and approaches to organize students who either graduated or dropped out of al-Aqsa University. As in Asif et al. (2014), the two data mining techniques (decision trees and Naive Bayes) were deployed. Both performed well, with high accuracy rates of 98.14% and 96.86%, respectively.

Guarín *et al.* (2015) focused on building a predictive model for student attrition (loss of academic status) at the Universidad Nacional de Colombia. Predictive modelling is a statistical

process commonly used for the prediction of future behaviour. The built-up models are used to describe how data is mined. While data mining shall be based on algorithms for the extraction and analysis of useful information through automatic discovery of hidden patterns, predictive models will help provide the next step in the analysis process. Two classification techniques—"Naive Bayes" and "decision trees"—were used. The data collected for the project were extracted from the Director of Admissions and Academic Information System and were restricted to only two semesters for students enroled in just two academic programs. The authors focused on two experiments: one to build a model that predicted students who would lose their academic status (dismissed students) and another for predicting the semester in which dismissal would occur. This work used a predictive model based on general academic performance, so there is still a need for building models based on the course level. Another point is the loss of academic status due to non-academic reasons that were not addressed.

Strecht *et al.* (2015) used classification and regression algorithms to predict student performance. The authors assessed students' achievement /failure by using classification techniques and used the multiple linear regression model to predict and analyse students' final grades. Although the classification and analysis presented better results compared with the regression, in my opinion, a precise summary of the statistical results is missing. It is not clear how the authors assessed the performance of the model. Further analysis is needed. The models were built based on specific non-academic parameters without considering other features that might influence student performance.

The other outline of the study/research was the comparison of the performance of different classifiers with the help of educational data mining, this was used in the various studies. Guarín, Guzmán, and González, (2015) aimed at predicting the enrolment of the students with the help admissions data. The scholars used applicants' data from the same concerned university. Many

classification models for learners were built. Results of various learners were compared, and the rules were identified from the j48 (an algorithm that used to generate a decision tree) and Rido (one of the rule induction algorithms) so as they can have the best.

2.5 Theme 2: Related work on Factors used to predict the performance

The above discussion showed that some models were built using academic attributes to predict students' performance, while others focused on non-academic factors. The following discussion will discuss what the main factors affected students' academic performance during their educational journey.

Yorke et al. (2005) addressed the extent to which the demographic background of students might affect their performance. Most of the previous work focused on academic backgrounds. Prior research focused mainly on student coursework—assignments, exams, etc. Vandamme et al. (2007) focused on first-year students, whereas Asif et al. (2014) produced a study focused on fourth-year students to predict their performance based on their historical data for the first and second year, and pre-university marks as an extra step. Another step was taken in this work that involved courses in which students indicated good/poor performance. This work is minimal and was focused on only data of the four cohorts from only two academic programs at the University of Pakistan were used.

Zafra et al. (2011) developed a new approach for using the online learning management system Moodle to predict student performance in individual courses based on three assessment tools (assignments, quizzes, and forums) available in Moodle. A comparison of traditional learning with multiple-instance learning (MIL) was made. Multiple Instance Learning is one of the

supervised learning types that deals with training instances that labelled in bags instead of considering individual labels. WEKA was a useful data mining open source for this project. Different data mining techniques were applied, and it was concluded that:

- The performance is significantly better when the multiple-instance learning approach is used compared with single-instance, and
- There is a need to improve the model to predict students' grades instead of predicting if she/he will pass/fail the course.
-

Osmanbegović and Suljić (2012) produced research mainly on applying three data mining techniques for predicting student performance at the University of Tuzla. This work was based on data collected from surveys conducted from 2010 to 2011 for first-year students. Twelve variables were tested to identify which ones played a central role in students' performance and which data mining technique(s) were more effective in achieving this goal:

- Naive Bayes performed better than decision trees and neural networks.
- There is a need for future research to expand on the built model by including more variables to make better predictions of student performance and identify techniques to include the data collection process within data mining tools.

AlShammary *et al.* (2013) analysed several research papers that addressed the effectiveness of applying Educational Data Mining techniques to learning outcomes. The authors concluded that data mining plays a significant role in predicting learning outcomes. The authors have pointed out some limitations of this work as there is still a need for future research work that addresses the following areas:

- Mining tools that could be used to improve teacher performance
- The effectiveness of data mining tools to improve learning outcomes

- Which data mining open sources are more effective?
- The effectiveness of professional development programs used to improve teacher performance

Ismail, (2015), built a predictive model to identify students who may graduate with a low GPA at Sultan Qaboos University, in Oman. The experiment focused on newly admitted students' performance. Fuzzy clustering data mining techniques were employed based on the related work reviewed by the author. It was concluded that most data mining techniques employed in the educational context aimed to 1) assess student performance and provide recommendations to the learners and 2) evaluate learning materials and identify certain students' learning behaviour. There is still a need to adopt fuzzy clustering methodologies to extract knowledge. For this paper, the authors used a clustering method known as "kEFCM: k-NN," which is based on the evolving fuzzy clustering method. This is an enhancement of the kNN clustering technique. The hybrid method's simulation results suggest that it generally outperforms K-NN notably when the location error is not more than 2 m.

The findings of this work are as follows:

- The accuracy of the educational data mining (EDM) approach depends on the size of the data set. The larger the data set, the higher the accuracy of the results, and
- Data mining techniques can accurately predict the accumulative GPA (AGPA).

Although the models were built successfully predicted the students' graduation AGPA, all the work was done based on high school results for newly admitted students, and the sample was 900 students.

Papamitsiou and Economides (2014) implemented learning analytics (LA) and linked it with the data mining techniques to highlight how both could improve the learning process. The authors defined LA as the area of research related to business in intelligence, web analytics, academic analytics, and predictive analytics. The review covered the literature from 2008 to 2013 and found that most studies focused on using data mining techniques to predict student performance based on factors such as grades in prerequisite courses that the student finished, assessment quizzes, and final exams, as well as student participation in certain activities.

Walldén et al. (2014) used a different approach to identify student performance by extracting students' data from Moodle as a learning management system, particularly students' data related to time, a user (student), types of resources used, the action performed, and name of the funds. The information extracted from Moodle can be used to enhance teaching and learning processes, as the instructor will be allowed to identify how actively his/her students used the teaching resources. The extracted pattern could help students help themselves.

A review was done by Shahiri (2015) to identify the main attributes that affect the performance in Malaysian institutions. The author has pointed out that yet, the research work in this area is still insufficient. Thakkar (2015) presented a survey that used data mining techniques from 2002 to 2014. He stated that there is a future need for research that focuses on non-academic attributes such as student behaviour, skills, and attitudes, and suitable data mining techniques are needed to measure the predicted academic performance (Aljahani, 2016). The review is easy to read and covered vital critical points that very few research works included. It also opened the door for a new gap that needs to be addressed: the need for integrated data mining techniques that address non-academic factors in predicting student performance, and not just their academic background.

Meier et al. (2016) developed an algorithm to predict students' final grades in a timely and personalized way based on their early assessments on assignments and quizzes. The developed model helped predict if a student would do good or bad in the course. This work was done based on data available to the instructor according to his assessment, which could be considered an advantage compared to the previous research work discussed above. The reason behind this is that the instructor does not need access to the data, which might be a challenge for privacy reasons; the instructor can use the Moodle based on the data available to the instructor from his or her assessment of students via different tools in the course. However, teachers could have trouble understanding and implementing this work. An expert in engineering or computer science is needed, which presents a limitation to work being generalized and performed by academic instructors.

Kumar (2017) surveyed students' performance prediction to identify what are the main attributes that affect a student's performance, and witch data mining algorithms could be used. The review showed that the Grade Point Average (GPA) and the internal marks of students are the most critical attributes for the prediction. Kumar (2017) identified that the (GPA) is the main contributor, and the review showed that the classification techniques are frequently used and in particular, the decision tree and Naïve Bayes are highly used, and it was admitted that more research is needed in the same area.

2.6 Critical Discussions

Data mining plays a crucial role for educational institutions by helping educational stakeholders predict and make decisions regarding students' academic status. With an increase

in the rate of student dropouts in higher institutions, the traditional manual systems only use numerical values to store and retrieve students' information. This shortcoming of this approach leads to the popularity of data mining to identify at-risk students, reducing the dropout rate and improving the retention rate. The traditional manual methods lack the accuracy, efficiency and artificial intelligence required for the analysis of data in the manner that data mining does. Data mining is an effective tool in predicting students' academic performances and identifying the main attributes that influence their performance. Although predicting students' performances can be measured through students' success, data mining provides a more effective way of finding the hidden patterns and providing suggestions in enhancing the students' performances. As data mining is effective in identifying the variables that affect students' performances, research continues to be undertaken to attain a greater understanding of the importance of data mining for students' educational improvements. The outcomes of the conferences reveal that data mining is a promising tool to improve students' learning as it is a useful tool in detecting and extracting relevant information from large volumes of data.

Several benefits are identified for data mining in predicting students' performances. Moreover, data mining has used the evaluation, planning, and counselling of students. Data mining also assists in enhancing the decision tree for students and the outcomes of the decision help in improving students' academic performances.

Kotsiantis et al. (2004) also use the machine learning to predict the students' performances accurately as the machine learning application is more effective in identifying low performing students and developing the strategy to help students who face academic risks. Moreover, data mining is a predictive model that could be used to improve graduate students' performance.

Although data mining appears to be effective in predicting students' behaviours in a small sample, there is a need to increase the sample size (Hung and Crooks, 2009). Moreover, matrix factorization and multi-regression techniques can also be used to forecast students' performances accurately and thereby reducing the error rate as compared with the traditional manual methods. However, the applications still need to be refined to overcome retention problems.

Kotsiantis, (2012) used classification algorithms for educational purposes for analysing the course completion rate and course preference. The decision tree algorithms were used to enrol students. It is also revealed that the association rules, clustering algorithms, and sequential patterns are used for predicting educational performances. Other points for discussions are different themes developed to enhance a greater understanding of machine learning for predicting students' performances. In theme 1, data mining techniques are identified as effective methods in predicting students' performances. The study identifies Fuzzy clustering data mining, kNN clustering technique, and enhanced kNN clustering technique, known as kEFCM: k-NN.

Márquez *et al.* (2015) believe that an increasing number of higher institutions are facing a retention problem. Although the traditional classification approach often partially solves this problem, classification algorithms are identified as a useful tool to improve retention rates. As the dropout rates are 7.9% in higher institutions, the feed forward neural network can be used to predict learning drop out through evaluations. Moreover, an artificial neural network is used to predict student retention rate in higher institutions with 87.2% accuracy. Furthermore, CAR is used to predict students' failure with 80% accuracy. The C4.5 decision tree is also used to predict freshman retention rate by 86.27% accuracy among university students (Márquez et al., 2015).

2.7 Gaps in the Literature

There are gaps in the literature that need to be investigated. The literature survey showed that most of the research were case studies that covered certain areas in an education context such as proposing a model to predict students' performance in a specific course. There is a need for research that will produce models that can be generalized to a wider population and on an institutional level, helping the higher educational institutions improve students' behaviours and skills, and thereby increase student retention.

This overview of the literature revealed the predominance of small datasets to predict student performances. Yehuala (2015) demonstrated the effectiveness of using large datasets to achieve accurate results. The author developed a model using WEKA software for a sample data of 11873 undergraduate students. The results were used to develop constructive and supportive decisions regarding educational systems.

Although this review focuses on students' performances, limited literature reviewed variables, such as quality of lectures, quality of teachers and school environment also affect student performances. Primarily, academic and non-academic variables play essential roles in assessing academic achievements. Academic variables entail all the variables which relate to them in class and school environment. The non-academic variables relate to the elements that affect the student away from a school or class environment. For example, students' standard of living, family background and ethnicity can also affect student performances.

2.8 Implications of the review and future direction

The present review showed that data mining techniques could play an essential role in educational systems by developing models to predict students' future academic performance. Based on the above discussion, different classification algorithms could be used to classify students according to their academic and non-academic attributes. Most of the research works reviewed focused on applying a decision tree, neural network as regression and clustering methods. Naïve Bayes and K Nearest Neighbor were used as classification tools.

Most of the data mining techniques were applied using WEKA as an open and free source, and it has performed well for this purpose. Most of the reviewed papers addressed the topic of small datasets and applied it on specific courses instead of on the overall performance. The focus was mainly on predicting students' performance based on certain historical academic attributes, whereas few papers considered the demographic information, and other non-academic attributes to test their impact on students' performance.

This research enhances the importance of data mining in predicting students' academic performances. This tool will assist school policymakers in identifying the risk factors that can affect student performances and develop policies to improve student learning. Essentially, institutions that record high rate of students' drop could face a risk of revenue and reputation risks. Thus, school management can use data mining to improve the learning process. The policymakers can also use the data mining tool to identify institutions that face a high risk of student drop out and formulate the policy to assist the institutions in improving the students' retention rates.

The findings of this review might help academic institutions, administrators, instructors, and researchers in using data mining tools to develop models to predict the future trends, students at risk (students with low academic performance), and help build an early warning system, as well as a recommender system for the courses in which the student needs to enrol.

Another critical issue is a need for further studies on the aspect of student retention and overall performance in Middle East higher education institutions, while paying attention to stored historical data to predict the future trends as few studies addressed this geographical area. This opens the door for future work on comprehensive studies that link a data mining approach with higher education systems in the Middle East to improve students' retention. The current study will address this area by applying data mining techniques on larger datasets of students enrolled in one of the national universities in the gulf countries.

2.9 Research Objectives and aims

The goal of this research is to develop a model to predict students' overall academic performance and identify the main attributes that influence students' performance. The plan is to address both academic and non-academic variables, an area neglected by current literature. The project aims to build a model that helps predict students' performance and their GPA (Grade Point Average) with the goal of early identification of students at academic risk in Middle East universities. This will allow university to employ an array of early intervention techniques with the result of increased student retention at Middle East universities.

2.10 The significance of the proposed work

The growth of higher education is a global phenomenon that has impacted the Middle East as well as other parts of the world. The implications of this growth are many, with student attrition being among the most prominent. Institutions should strike a balance between providing high quality educational experiences for their students and facilitating access and accessibility to a wide range of students. In pursuing this balance, institutions must invest in planning their retention efforts. Most of the time, planning means conducting research, looking carefully into data and making informed decisions based on data generated (Hagahmed, 2014).

This research enhances the importance of data mining in predicting students' academic performances. This tool will assist school policymakers in identifying the risk factors that can affect student performances and develop policies to improve student learning. Essentially, institutions that record high rate of students' drop could face a risk of revenue and reputation risks. Thus, school management can use data mining to improve the learning process. The policymakers can also use the data mining tool to identify institutions that face a high risk of student drop out and formulate the policy to assist the institutions in improving the students' retention rates.

2.11 Research Questions

The following research questions guided the study:

R_1 Which data science algorithm(s) is the most appropriate and effective in developing a predictive model to predict students' GPA?

R_2 What are the main factors that affect students' GPA in each academic year?

The present work will address the research questions by exploring the possibility of building predictive models that could be used for the following purposes:

- Predicting the student's Grade Point Average (GPA) for students enroll in one of the gulf countries university (more than 19000 students), and
- Identify the factors that affect student's performance.

This first study was conducted at one of the Gulf universities as part of Middle East higher education institutes and was built on discovering the knowledge from stored data in the Student Information System (SIS) instead of a general review of the student's transcript.

So, the research questions and objectives are summarized in the following matrix:

Table 2. 1Research objectives and questions

Research Objectives	Research Questions
To build a predictive model to predict students' performance based on their previous results, and grades achieved, as well as their academic history.	Which data science algorithm(s) is the most appropriate and effective in developing a predictive model to predict students' GPA?
Identify the main factors/ variables that affect students' performance and the relationship between a student's performances.	What are the main factors that affect students' GPA in each academic year?

Chapter 3: Research Methodology

3.1. Introduction

The chapter presents the research design and methodology. This includes a description of the sample; data collection procedures, including sampling methods, protections for human subjects and ethical approval and data analysis procedures.

Open source software was used to build predictive models by using data science algorithms. An evaluation of the models is detailed. In order to answer the research questions, different Machine Learning algorithms were applied by using Rattle which is a free graphical user interface (GUI) builder for data mining with R. According to Williams et al. (2011), Rattle started out using the Python (1989) programming language then, soon moved to R directly.
The following Machine Learning (ML) algorithms were applied and compared to identify which algorithm(s) works better for the model. These are:

- Multiple Linear Regression to predict the numerical response variable (students' GPA)
- Logistic Regression to predict the academic status (Binary response variable for a specific category)
- Regression Trees
- Random Forests

A comparison between the above algorithms has been discussed to select the classifier/ predictor. Furthermore, A summary description of each of the above algorithms is presented in the coming sections.

3.2. Author relationship to the study

The author of this research is serving as a faculty member in the same university subject to this study. Being a faculty in the same academic institute helped and strengthens the requested access to the dataset needed for this research work after granted the official approval by the university research ethics committee. Furthermore, the author's work in the same institute helped in understanding students' levels, characteristics, academic programs, different levels of students' academic standing (Good standing, academic probation, academic dismissal, etc.). Despite the knowledge and familiarly of the academic institute that the author has over the past years of experience but the data collected objectively and through official channel starting from submitting the official request to get access to the data until getting access to the data by Student Information System (SIS).

3.3 Research Design

This study employed an experimental design and considered the main phases of building, testing, evaluating, and deploying the model. Data mining techniques were used to build predictive models to answer the research questions formulated in the previous chapter. Rattle and R were used as they are free open sources that provide a sophisticated environment for data analysis, visualization, and statistical analysis.

The concerned university accepts students under two different categories. These are:

FN: **Foundation program**: Pre-College program (Mandatory for all Science, Engineering, Health, Medicine as well as Pharmacy students) where students need to pass specific mathematics, and English Language courses to be qualified to move the College/ Major.

The required courses in the Foundation program are pass/ fail the class and students don't earn credit hours or GPA unless the student exempted from some of the Foundation Program

Courses, then they are eligible to take specific courses with credit hours while they are in the Foundation program and can get a GPA. So, some students in the Foundation program have GPA, while the majority do not.

UG: **Undergraduate:** students did not need to pass the Foundation program and were accepted directly into the College. These students under Arts, Humanities, Social Sciences, Business, Law, Sharia, and Islamic studies, etc. streams.

As the intake at the university happen twice a year in each term, so the dataset will be divided according to the admission term as follows:

Term 1: admitted in fall

Term 2: admitted in spring

For each cycle, students will be grouped based on their levels, and the numbers of terms spent at the university.

The research is designed to compare like groups with like groups. Each group requires different courses and students in the Foundation Program need to pass classes without GPA while the UG Students are required to take classes with credit hours and will earn GPA by the end of each term. Therefore, students were split into two groups, and the models will be built based on the structure of each stream.

The research is designed as follows:

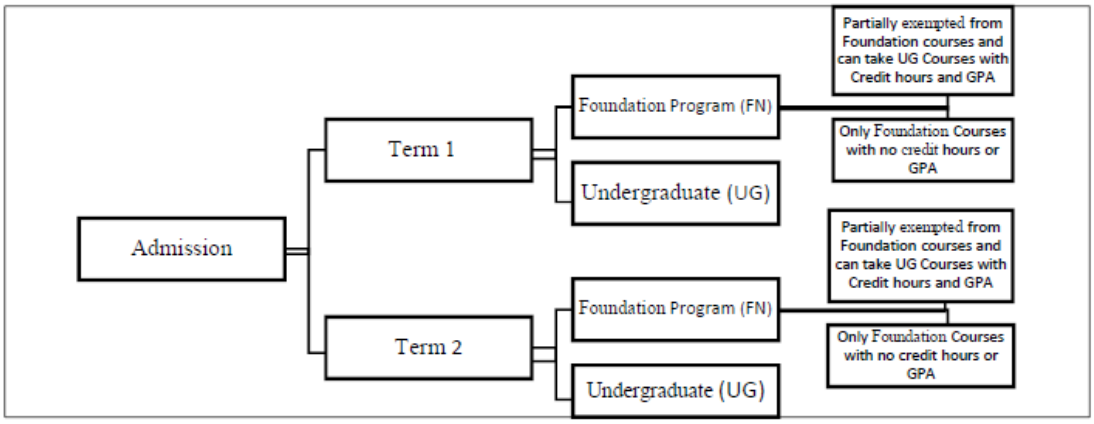


Figure 3. 1 Research flow design

For Foundation students, the models will be built based on the following classification:

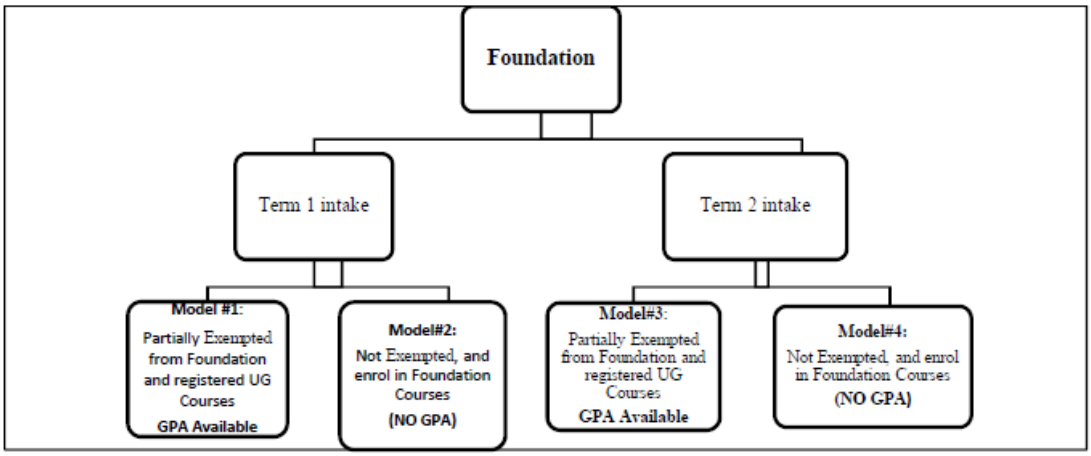


Figure 3. 2 Research flow - Foundation Program

For the first year and after two terms, the models will be developed for each group as follows:

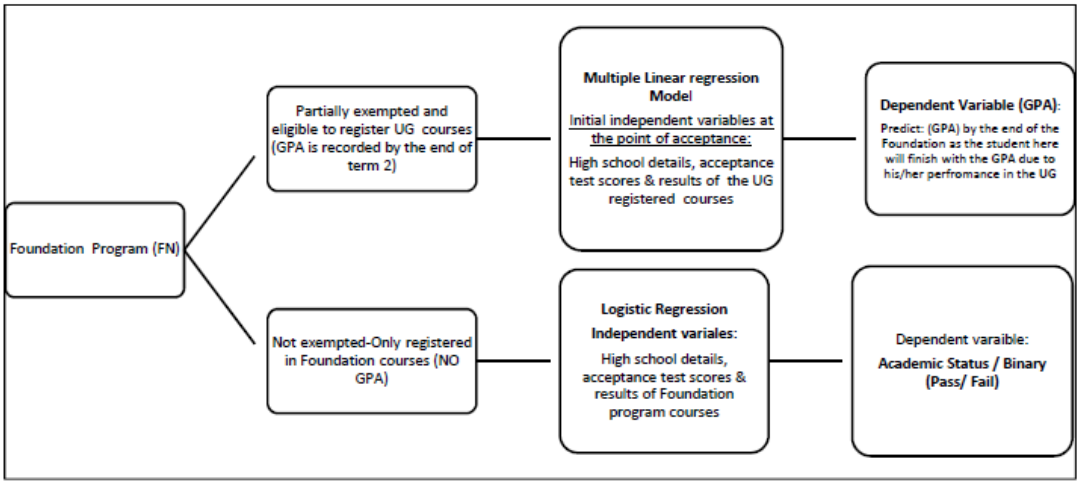
Group 1: Foundation Program

Students who enrol in the Foundation Program are required to pass specific Mathematics and English courses. It is a two-term program, and all classes are pass/fail without GPA.

For this category, the independent variables are students' results at high school, admission tests, as well as the results of the pass/, fail courses at the Foundation level as a pre-College program.

So, the dependent variable will be the academic status (pass/ Fail) by the end of the Foundation year. For those who are partially exempted from the Foundation and registered UG courses while they are in the Foundation, they will finish the Foundation program with GPA and hence, for this category the dependent variable will be the GPA of term_2.

The following process will be applied for Term 1 and Term 2 separately.



Group 2: Undergraduate students

Data for Undergraduate (UG) students were split based on the term admitted and models were built for each cohort separately. GPA is not static and may change with each term. Since the goal of this research is to build a predictive model for academic jeopardy, it is imperative to track GPA over time. If the student earned a GPA less than 2.00/ 4.00, then they will be placed on academic probation. Dismissing students from the university is not one step. If the GPA became low for certain number of terms, then the student is placed on final academic Probation. If the student failed to raise the GPA, then they will be dismissed from the university. In order to provide the academic support needed by students, it is important to explore the possibility of predicting the term GPA so an immediate academic support could be provided to the student before getting a final probation or dismissed from the university. Splitting the data by term

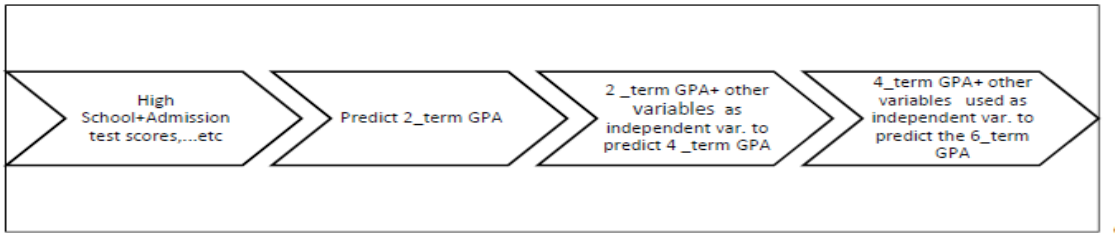
allowed for building modes to predict the GPA for students who have a 2_term GPA, 4_term GPA, 6_term GPA, etc.

The independent variables for the two-term group includes: High school, gender, school region, nationality, admission tests, etc. The dependent variable is end of 2nd term GPA.

In order to build a model to predict 4 term_GPA (the GPA by the end of year 2) the independent variables are:

- The GPA by the end of the previous year (year1)
- The number of failed courses, and
- All other variables (Earned hours, age, school origin…etc.)

The above process will continue to predict the GPA of students who have 6_term GPA and the outcomes of the previous year will be used as independent variables for the next model.



6th Figure 3. 4 GPA and the outcomes

3.4. Method adopted

3.4.1. Resources of the data

The concerned university has provided the dataset needed for this project. The decision made to select this university because it is the only national university in the country that serves most students, and it is the first option to most of them. The datasets include the List of undergraduate students from the period 2003- 2016.

A set of more than 19,000 student records have been extracted from the Student Information System (SIS) as well as the Information Technology (IT) Section. The table below shows a description of the datasets provided.

Table 3. 1 Description of the range of the variables

Variable	Students enrolled in Term 1	Students enrolled in Term 2
Admitted Term	students admitted in Fall	Students admitted in Spring
Age	**Range: 19- 39 years**	**17-55 years**
level	**UG** (Undergraduate students classified as undergraduate): **15457 students** **FN** (Foundation: Students admitted in the Pre-College Program: Foundation Program): **615 students**	UG: 3236 FN: 151
Earned hours	The number of credit hours that the student earned. Minimum Hours: 0 CH Maximum Hours: 173 CH	0-152 CH
Accu_GPA	The accumulative GPA based on the GPA of all previous terms. Minimum GPA (0.00): Maximum (4:00)	Minimum GPA (0.00) Maximum (4:00)
Term_ GPA	The GPA for each term. Minimum GPA (0.00): Maximum (4:00)	Minimum GPA (0.00): Maximum (4:00)
High School %	The admission decision mainly built based on the student's high school performance. Lower High school: **70%** Highest High school: **99%**	52%- 99%

Further details about the other variables are illustrated below.

Table 3. 2 Further details about the dataset variables

Variable	**Description**
Accu_GPA	The accumulative GPA based on the GPA of all previous terms. Minimum GPA (0), Maximum (4)
Failed Courses	No. of courses that the student failed.
High School %	The admission decision mainly built based on the student's high school performance.
APL_Accuplacer	A placement test that the university conduct for all newly admitted students.

APIC_Integ_Core	Foundation course offered for students who join the Foundation program.
APLU_Lang_Use	English Foundation course offered for students who join the Foundation program.
APWS_Writing_Workshop	
APRS_Reading_Skills	
APSM_Sentence_Meaning	
APLG_Llistening	
TOEFL	English International Language test.
IELTS	
APLA_ Arithmetic	Math. Foundation course offered for students to join the Foundation program.
APCL_COLL_Level_MATH	
SAT	Global test score
ACT	
IC3	
School Origin	This variable classifies student based on high school. Either local (Qatari) or international school.

In addition to the detailed dataset, comprehensive policies and procedures about academic standing, academic probation, and dismissal policies were required. This information was extracted from the university catalogue, and published policies and procedures.

3.4.2 Ethical Approval and Protection of Human Subjects

In a paper presented at the 8th International Conference of Educational Data Mining, Sabourin *et al.* (2015) raised the critical issue of confidentiality of student data and the challenge of using it in data mining research. It is recommended that the researchers maintain transparency while mining data. They should also preserve accountability for any potential breach of privacy. There is a need for pro-activeness in the implementation of confidentiality to ensure that data is not used for the purpose for which it was not intended and that it is secure. Based on this, and to get the ethical approval from the university provided the dataset needed for this research, an official request submitted to the University Institutional Review Board (IRB) for ethical

approval. This is one of the university's standing committees which support the academic research.

The request to release the datasets and needed information presented to the committee including all supporting documents. The Institutional Review Board (IRB) issued the official approval to release the datasets on September 4th, 2016. A copy of the mentioned ethical approval is attached in appendix A.

This step followed by another step to issue the ethical approval letter from the University of East London (UEL). The author went through all official channels and steps according to the UEL policies and procedure. The official ethical approval for the collection and storage security and sensitive information issued by UEL on April 6th, 2017 under the reference number UREC161745. A copy of the approval is attached in Appendix A.

Information in the database containing the unique participant identification numbers will remain for three years after the completion of the study and then be destroyed. The database for the study will be password-protected and secured, only accessible to the researcher. Three years after the completion of the study, all data collected for the study will be destroyed or deleted from the researcher password- protected hard drive.

3.4.3. Data analysis

Several data science techniques were applied such as classification (Multiple linear Regression, Logistic Regression, regression trees, and Random Forest) based on historical data available. Quantitative data analysis and evaluation approach was carried out on datasets including

numerical variables to compare all algorithms results. The process started with splitting the dataset based on the admission term (Fall/ Spring). Admission is competitive and based on an annual admission capacity, so usually, there is a significant number of students who join the university in term 2. There are two reasons behind this. One of the reasons is related to the colleges' capacity as some students might don't get a chance to join in fall term due to the admission capacity. The other reason is many (male) students prefer to postpone their admission and start full-time work.

The box plots below illustrate the high school percent for students admitted in each term.

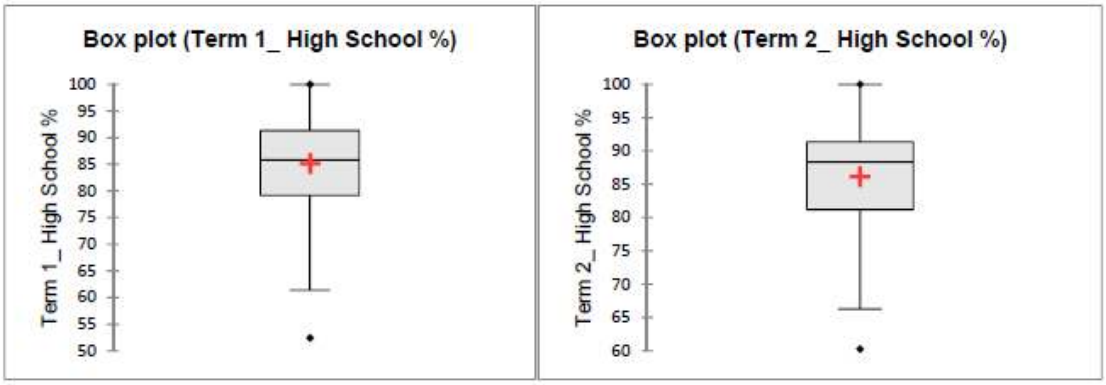


Figure 3. 5 Boxplot for the high school % of students admitted in term 1 &2

In terms of partitioning the dataset, 70-75% of the data to train the model, and the remaining percent used to test the model. For the model's evaluation, the following approach was used during the evaluation process.

In applying Multiple Linear Regression models, the author used three statistics to assess the model's performance. These were:

- R- Square indicates the goodness of fit of the model

- F- Test indicates the null hypothesis. It shows that whether the relationship between the response variable and the set of predictors is statistically significant or not.
- Root Mean Square Error (RMSE): the square root of the variance of the residuals. The lower the value, the better fit.

Logistic Regression model was used to predict students' academic status (pass/fail) on the Foundation Program level to assess the model performance, using the observed and predicted diagnostic table which includes the answer of true cases (observed cases that are predicted to (True Positive " TP) as well as the number of True Negative (TN). Furthermore, the Receiver Operating Characteristic (ROC) which present the plot of the sensitivity (the proportion of true positive) vs. (1- Spesivity (proposition of negative cases)). ROC curve assesses the overall how well the model predicts the target variable. In addition, the probability of Chi-Square is given an indication about the influence of the variable in the model. It is equivalent to the Fisher's F test. For the Regression Tree, the Mean Absolute Error was used to measure the performance of the model.

3.5. Conclusion

This chapter presents the methodology used to apply data mining techniques to predict students' GPA. The dataset was described and protections for human subjects were delineated. The dataset was split by term (2, 4 or 6) and program (Foundation or Undergraduate) and this process was detailed. Various methods of analysis were applied to the different groupings of students. Next chapter will discuss different types of applied data science techniques and how they used on the datasets.

Chapter 4 Machine Learning

4.1. Introduction

Machine learning, a subfield of artificial intelligence, is a tool for turning information into knowledge. The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. In the past 50 years, there has been an explosion of data. This mass of data is useless unless it is analysed to find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that would otherwise be difficult to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making (Edwards, 2018).

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with \machine learning methodologies or analyze the impact of machine learning processes (Tagliaferri, 2017).

In this chapter the author reviews the common machine learning methods of supervised learning, and common algorithmic approaches in machine learning, including the linear/ logistic regression algorithm, regression trees, and Random Forest.

4.2 Datasets

As discussed in the previous chapter, the datasets provided by the academic institute includes two different categories. These are:

- **Foundation (FN) program**: Students enrol in the Pre-College Program and students' academic status will be either Pass or Fail by the end of the program without getting numeric GPA, and
- **Undergraduate (UG):** Students either passed the Foundation program or it is not part of their degree requirements and declared their majors (or eligible to declare their major) in one of the academic programs available at the university.

  This chapter starts with a brief discussion about the Machine Learning algorithms, and the implementation phase on each category of students to either predict students' academic status, pass / fail for students enrolled in the Foundation program, or to predict the numerical GPA for undergraduate students.

4.3 Data Mining and Machine learning

Data mining refers to extracting knowledge from a large amount of data. Data mining is the process to discover various types of patterns that are inherited in the data and which are accurate, new and useful. Data mining is the subset of business analytics; it is similar to experimental research. The origins of data mining are databases, statistics. Machine learning involves an algorithm that improves automatically through experience based on data. Machine learning is a way to discover a new algorithm from the experience. Machine learning involves the study of algorithms that can extract information automatically. Machine-learning uses data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future outcomes. For Reference https://www.educba.com/data-mining-vs-machine-learning/.

Since both concepts deal with data, and it looks like that both are the same. Although as indicated by Lantz (2015), one of the potential points of distinction is that Machine Learning focuses on teaching the computer to use data to solve a problem, while data mining is showing the network to extract the knowledge from hidden information and to identify patterns that human use to solve a problem.

Machine learning techniques are divided into two types. These are:

- **Supervised machine learning algorithms:** in this type of algorithms predictive models are used to predict specific targets based on other values in the dataset and can apply what has been learned in the past to new data using labelled examples to predict future events. In this type of model, explicit instruction is given about what the model needs to learn. An optimization function (the model) is formed to find the target output based on specific input values. Starting from the analysis of a known training dataset, the learning algorithm

produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- **Unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data. According to Lantz (2015), there is no single feature or target value to learn, it is all about pattern discovery, which used to identify the associations within the dataset values.

Since the main purpose of this work is to predict students' GPA, the focus is on the first type of Machine Learning which is supervised learning by using different algorithms to explore the possibility of building predictive models to predict the student's future GPA based on certain historical values. Four supervised learning algorithms used and compared to find out the reasonable method to predict the student's GPA. These are:

- Logistic Regression
- Multiple Linear Regression
- Regression Trees
- Random Forest.

4.4 Supervised methods

4.4.1 Logistic and Linear Regression

According to Lantz (2015), logistic regression is used to model binary categorical outcomes. O.D.& P.A (2017) presented multiple regression models to predict the student's academic performance (SAP) using data of the Department of Computer Science in Nigeria. In this type of regression, the regression equation models data by using similar slope- intercept format as y = a+ bx. In this case, the machine's job is to find the best values of a and b, so the line is best able to fit and relate to the proved x values to the value of the target variable.

Linear regression is used for not just classification, but also for predicting a numerical value and estimating the relationship between these numerical values. This relationship is between the target (output) variable, and more than one independent variable or predictors. In case more than one independent variable used, it is called multiple linear regression which is the case in the present work.

4.4.2 Regression Trees

Decision trees are one of the supervised learning algorithms that used widely for data exploration purposes. According to Yadav et al. (2012), a decision tree is a flow- chart like a tree structure. Rectangles denote each interval node, and ovals indicate each leaf nodes. There are different two main types of decision trees. These are:

- Classification tree: used if the response (target) variable is categorical, and
- Regression tree: used if the target variable is a real numerical number as it is in our case in the present work (Student's GPA).

There are different types of decision trees. Some of the conventional and widely used methods are:

CART: used for Classification and regression tree.

CHID: (CHI- squared Automatic Interaction Detector). In this case, multi-levels are performed.

Koksiantis & S.B (2012) presented a decision support system for predicting student's grades. The author pointed out that although the model trees are small and more accurate than regression trees, as indicated by Wang & Witten (1997), regression trees are comprehensible.

Ibrahim & Rusli (2007) compared the machine learning algorithms (linear regression, Artificial Neural Network (ANN), and decision tree) to predict the performance of 206 undergraduate students. The result showed that the ANN model outperforms than the other two algorithms. This study was restricted on relatively small size dataset.

4.4.3 Random Forest

Random forest is one of the supervised machine learning algorithms. It is considered as a collection of decision trees. It is an ensemble learning technique for both classification and regression tasks, which make it widely used. Furthermore, one of the main advantages of random forest algorithms is that it is easy to measure the importance of the model's features or variables by computing the score automatically for each node of the tree.

Compared with a decision tree, in the decision tree, a set of rules are generated based on the features that describe the tree, while in random forest algorithm, observations are selected randomly to build several decision trees, and then takes the average of the results.

4.5 Summary

The present discussion provides a brief about the Machine Learning techniques that will be used in modelling students' dataset to build predictive models to predict the GPA for each cohort.

Many scholars (Romero & Ventura (2010), Tair and El Halees (2012), Natek and Zwilling (2014), Vandamme *et al.* (2007), Tsai *et al.* (2011) and Veeramuthu *et al.* (2014)) have carried out research concerning the prediction and assessment of student's results and performance in various universities. In Iqbal et al. (2016), an analysis of different international studies and examination of the admission criterion of Qatar University was done, this was also intended to establish the exact factors that can be used in the prediction of the students results (GPA) while in their first year of study at Qatar university. With reference to the attained results, it was discovered that the high secondary certificate performance and the entry test are the key factors in the prediction of the student's performance in the first year of study. The study is further broadened to include a research on the examination of the effectiveness of the student performance in Qatar University in various courses they are enrolled in.

In the educational data mining arena, there is a recent surge in publishing and research papers. Take an example, a classification model to be used in the prediction of the best study track for the students in school. The data is gathered from six schools in Jordan, it adds up to 248 instances. The decision tree had an accuracy of 87%.

Retention of students in universities is serious matter of concern that must be addressed. Limited academic support and lack of appropriate academic advising are among the leading

causes of undergraduate dropouts in universities, starting as early as first year. On that note, the first year of undergraduate study is largely known as the break or make year. The lack of support on the complexity of the course and the course domain can easily lead to a student's demotivation and consequently withdraw from the course. It is therefore important to help students as early as possible if they are to survive in these higher institutions of learning.

One of the solutions that can be used to help these students is the early grade prediction system. In this system, the academic advisors are able to predict the results of a given student and advise him/her accordingly. This also improves on the motivation of the students and thus retention in the institution. In addition, the use of machine learning coupled with educational data mining (EDM) is crucial in the learning of students in such institutions. Many other models can be designed to help in the prediction of students' grades in the various enrolled courses and this will be helpful in the provision of information that will subsequently lead to retention of students in these higher institutions. The information gathered can also be helpful regarding identification of students whose chances of attaining low grades are high. In this way they will be given special attention to help them improve on their grades (Iraji et al., 2012). The next Chapter will present the implementation process for each of the above methods and the outcomes.

Chapter 5 Implementation of Machine Learning Algorithms for students enrolled in the Foundation Program

5.1 Introduction

This chapter presents the implementation process of the Machine Learning (ML) algorithms discussed in the previous section. As mentioned, students are admitted twice a year in each term and classified into two categories based on their academic level at the admission period. For example, all Science, engineering, Pharmacy, and Medicine students need to pass a pre-college program called "Foundation Program" before they are fully admitted as Undergraduate (UG) students. That must also be happen before they declare majors, whereas, students admitted in other colleges (Arts, Business and Economic, Islamic Studies, Law, and Education) are not required to pass the Foundation program, and they admitted to the respected college immediately.

All Foundation courses are graded pass/fail and do not earn credit hours. In other words, students finish the Foundation program without earning a GPA, while for other colleges, students can register for undergraduate (UG) courses, which are letter graded, and hence they receive a GPA by the end of each term. Based on this distribution, the implementation process will be divided into different phases to cover each category individually**.**

Then the implementation process will consider each of the two admission terms individually. A comparison between term1 and term 2 admissions for all levels will be addressed to identify any significant differences in the built models, and factors that affect the performance either for each group: 1) admitted term 1 and foundation; 2) admitted term 2 and foundation; 3) admitted term 1 and UG; 4) admitted term 2 and UG. The next sections present the implementation process of the Machine learning algorithms for each category.

5.2 Foundation Program

5.2.1 Distribution of the Foundation Program students

Students admitted in the Foundation program can be classified into two groups or types, in addition to the term in which they were admitted. Students admitted in the Foundation program with international test scores such as IELTS, TOEFL, SAT, ACT, etc., are eligible to be exempted from some of the Foundation courses and be allowed to register undergraduate courses with letter grades. This type of students will finish the Program with a GPA that relates to the designated undergraduate courses. Students without international test scores must pass all the Foundation Program courses (Pass/ Fail) before taking undergraduate courses, so complete two terms without earning a GPA. The present work will address each group individually, and separately as the requirements and the target for each group are not the same. The diagram below illustrates the distribution of the Foundation program students based on the above discussion.

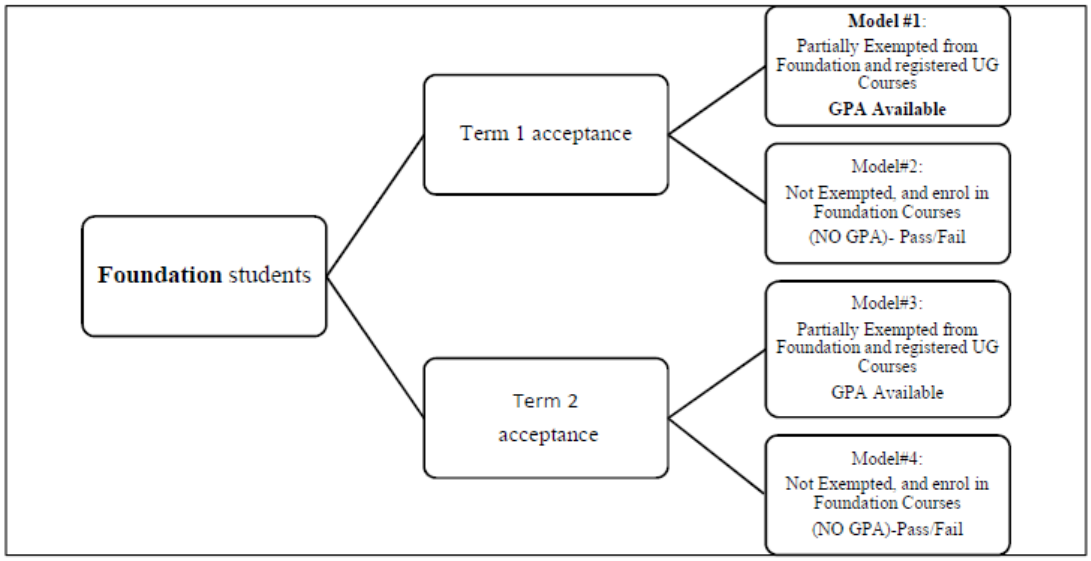


Figure 5.1 Distribution of the Foundation Program students

The next discussion will address the two categories of Term1, and models built for each category. The table below illustrates the size of the datasets of the Foundation Program for the two terms.

Table 5. 1 Size of the datasets of the Foundation Program for the two terms

Term	Not exempted from the Foundation program	Partially exempted and allowed to register (UG) Courses	Total
Term 1	596	19	615
Term2	140	14	154
Total	736	33	769

As we see from the above frequencies, the number of students who are partially exempted in both terms is small, so we will pay attention to the datasets of students who are not exempted and enrolled in the Foundation Program.

5.2.2 Term 1 acceptance: Students Not exempted from the Foundation Program

The University with an official approval by the University Research Ethics committee provided the dataset of all Foundation program students. It includes details about students' age, nationality, gender, high school results, as well as the performance in the Foundation program (Math, English) courses. The table below lists the variables included in the datasets of this category as follows:

Table 5. 2 Variables included in the datasets

Variable	Description
Age	Student's age
Gender	Student's gender
High School %	High school result (%)
APLU_Lang_Use	English course the Foundation
APIC_Integ_Core	English course the Foundation
APEA_ELEM_ALGEBRA	Math. Course at the Foundation
IELTS	International English Test
APRS_Reading_Skills	English reading course the Foundation
APWS_Writing_Workshop	English writing course the Foundation
APLG_listening	English listening course the Foundation

APSM_Sentence_Meaning	English course the Foundation
APLA_ Arithmetic	Math. Course at the Foundation
APCL_COLL_Level_MATH	Math. Course at the Foundation
ACT	International Test score

Since all the above Foundation courses are Pass/ Fail without numeric GPA, and since the goal is to predict student performance by the end of the Foundation program, then the final academic status (response variable) in this case is a binary variable. So, logistic Regression is carried out to predict the academic status (pass/fail) of the Foundation students.

5.2.3 Factor Analysis and Logistic Regression

We used a Varimax rotation to extract the most important of the 13 values. The results showed that only two factors could be extracted and a representative variable with highest loading scores were extracted as illustrated below.

Table 5. 3 Kaiser-Meyer-Olkin measure of sampling adequacy

Variable	KMO Score
Age	0.749
High School %	0.797
APIC_Integ_Core	0.556
APWS_Writing_Workshop	0.515
APRS_Reading_Skills	0.500
APSM_Sentence_Meaning	0.552
APLG_listening	0.478
APLA_ Arithmetic	0.804
APLU_Lang_Use	0.563
APEA_ELEM_ALGEBRA	0.773
APCL_COLL_Level_MATH	0.635
IELTS	0.569
No of Failed courses	0.821
KMO	0.561

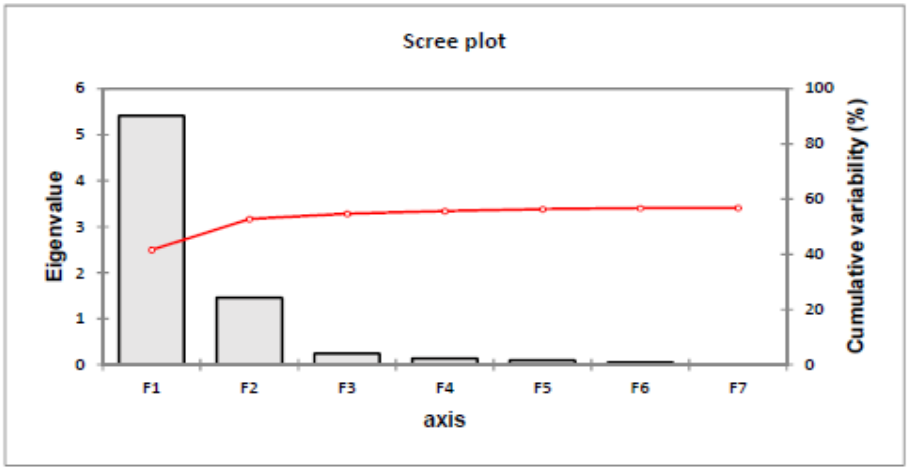


Figure 5. 2 Scree Plot

Based on the above outputs, we can see that two factors could be selected with 53% variability in the dependent variable. For these two factors, and according to individual KMO scores > 0.5 as well as the factor loading scores after the varimax rotation, then two representative variables were extracted to be included in the model. These are:

- APIC_Integ_Core: which is an English language course offered at the Foundation level
- APEA_ELEM_ALGEBRA: which is an Elementary Algebra course offered at the Foundation level.

The Logistic regression model has been carried out to predict the academic status of students enrol in the Foundation program. 67 % of students are female and the 33% are male. About 88% classified as “Not Fail”, and the remaining 12% are failed. The discussion below illustrates the outputs we reached after carrying out the models as follows:

Table 5. 4 Type II analysis (Variable Status)

Source	DF	Chi-square (Wald)	Pr > Wald	Chi-square (LR)	Pr > LR
APIC_Integ_Core	1	22.051	< 0.0001	23.849	< 0.0001
APEA_ELEM_ALGEBRA	1	10.182	0.001	11.429	0.001

Based on the above two table, and from Type II error and by looking at the probability of the Chi-squares that the variable most influencing is APIC_Integ_Core.

The equation of the model (Variable Status) is:

Pred(Status) = 1 / (1 + exp(-(-1.39159769745891+9.47501818030457E-03*APIC_Integ_Core+1.93640132520604E-02*APEA_ELEM_ALGEBRA)))

As we see from the above table the probability of Chi- Square is less than 0.0001, then this is an indication that the variable is brought significant information.

The confusion matrix has been formed and illustrated below. The matrix below indicating the performance of the classifier, and it shows the percentage of students that well classified (the number of students that well classified/ the total number of students). As we see from the matrix below, 88.24% of instances are correctly classified.

Table 5. 5 Confusion Matrix

Actual				
Predicted	Fail	Not Fail	Total	% correct
Fail	0	70	70	0.00%
Not Fail	0	525	525	100.00%
Total	0	595	595	88.24%

Furthermore, the ROC curve is used to evaluate the performance of the model by means of the area under the curve (AUC) as illustrated below.

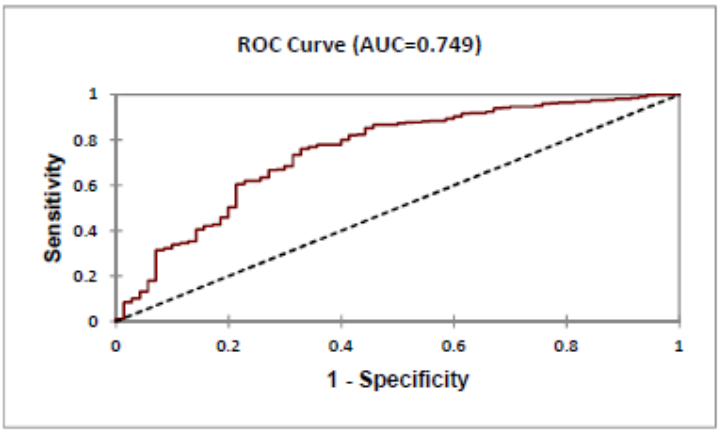


Figure 5. 3 ROC Curve (AUC=0.749)

5.2.4. Term 1: Partially exempted from the Foundation Program

The size of the dataset of this category is small (n=19). Since the number of this group is too small, this category was not addressed.

5.2.5. Term 2: Not exempted from the Foundation Program

In general, the number of students admitted in term 2 (spring) is smaller than those admitted in term 1. Most students apply to join the university in term 1 (fall), and according to the university annual capacity, there should be number of students who meet the admission requirements, but they didn't get accepted due to the colleges' capacity. These students usually advised to apply again and join the university in term 2 when there are fewer applicants.

The dataset provided by the University includes 140 students who were not exempted from the Foundation program and needed to fully enrol in the Foundation's courses without exemption. As pointed out before, these students need to pass Foundation (Math / English) courses before they may take undergraduate courses for credit.

As discussed above, the target variable here is the binary variable, academic status. Therefore, Logistic regression was used to predict the final academic status by the end of the Foundation program. The independent variables are all Foundation courses as well as high school results, age, gender, number of earned hours, etc. 84% of students in this group were in local schools, while the remaining 16% are from private (international) schools.

The age range is 18-39 years as some students who apply to join in term 2 didn't join the University immediately after they finished high school. They prefer to undertake a full-time job, and then return to the university to get the degree. This is a common practice, especially for male students.

The correlation matrix was formed to check the internal correlation in order to identify if all 14 variables should be included in the model. The result showed high correlation between some of the Foundation Program courses as highlighted in yellow in below matrix.

Table 5. 6 Correlation Matrix

Variables	APCL_COLL	Age	High School	APEA_ELEM_ALGEBR	APLG_listening	APLU_Lang_Use	APIC_Integ_Core	APWS_Writing_Works	APSM_Sentence_Meaning	APRS_Reading_Skills	APLA_Arithmetic	IELTS
APCL_COLL_Level_MATH	1	-0.032	0.095	0.089	0.107	0.172	0.162	0.164	0.140	0.140	0.111	0.048
Age	-0.032	1	-0.320	-0.250	-0.207	-0.224	-0.214	-0.200	-0.160	-0.169	-0.277	-0.006
High School %	0.095	-0.320	1	0.584	0.222	0.348	0.311	0.323	0.273	0.317	0.288	0.036
APEA_ELEM_ALGEBRA	0.089	-0.250	0.584	1	0.177	0.403	0.331	0.352	0.278	0.358	0.348	0.091
APLG_listening	0.107	-0.207	0.222	0.177	1	0.765	0.857	0.816	0.783	0.667	0.146	0.033
APLU_Lang_Use	0.172	-0.224	0.348	0.403	0.765	1	0.921	0.939	0.801	0.786	0.340	0.081
APIC_Integ_Core	0.162	-0.214	0.311	0.331	0.857	0.921	1	0.972	0.922	0.891	0.322	0.133
APWS_Writing_Workshop	0.164	-0.200	0.323	0.352	0.816	0.939	0.972	1	0.958	0.819	0.318	0.147
APSM_Sentence_Meaning	0.140	-0.160	0.273	0.278	0.783	0.801	0.922	0.958	1	0.771	0.266	0.186
APRS_Reading_Skills	0.140	-0.169	0.317	0.358	0.667	0.786	0.891	0.819	0.771	1	0.348	0.130
APLA_Arithmetic	0.111	-0.277	0.288	0.348	0.146	0.340	0.322	0.318	0.266	0.348	1	0.139
IELTS	0.048	-0.006	0.036	0.091	0.033	0.081	0.133	0.147	0.186	0.130	0.139	1

To identify which variables, need to be included, factor analysis was performed with varimax rotation. eigen values greater than 1 considered. In terms of Kaiser-Meyer-Olkin (KMO) measure; All individual measurer scored 0.5 and above as well as the overall score considered.

After running the varimax rotation, the outputs are illustrated as coming below.

Table 5. 7 Kaiser-Meyer-Olkin measure

variable	KMO score
Age	0.738
APEA_ELEM_ALGEBRA	0.803
High School %	0.725
APIC_Integ_Core	0.569
APLU_Lang_Use	0.526
APWS_Writing_Workshop	0.547
APRS_Reading_Skills	0.537
APSM_Sentence_Meaning	0.516
APLG_listening	0.556
IELTS	0.619
APLA_ Arithmetic	0.824
APCL_COLL_Level_MATH	0.957
Overall KMO	0.565

For the factor scores, we considered all variables associated with each factor scored 0.5 and above, and from each group a representative variable with the highest score was selected to represent all variables within the same factor instead of including all variables in the model. As seen from the above graph, only five variables are contributing to the model. so, and as seen from the classification matrix, the model is well predicting 73% of the data.

Table 5. 8 Classification Matrix

	Actual			
Predicted	Fail	Not Fail	Total	% correct
Fail	46	17	63	73.02%
Not Fail	20	56	76	73.68%
Total	66	73	139	73.38%

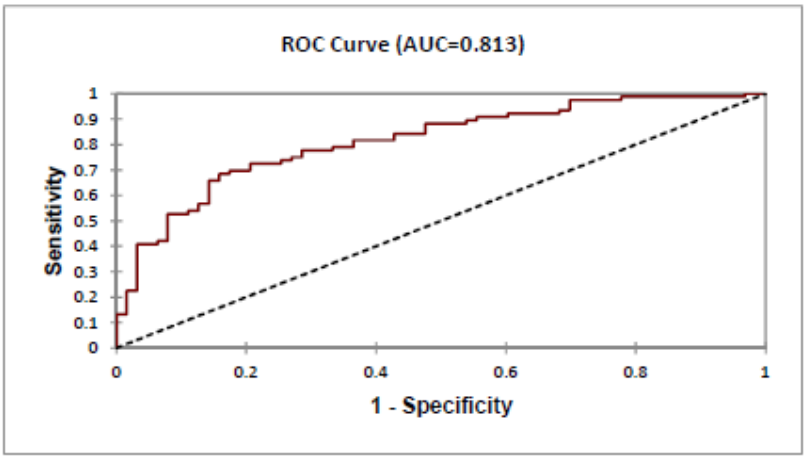


Figure 5. 4 ROC Curve

As we see, the Area Under Curve (AUC= 0.813) which is a good score.

5.2.6. Term 2: Partially exempted from the Foundation Program

The last group of the Foundation students are those who entered the University in term 2 and were partially exempted from the Foundations courses. These students were allowed to register for undergraduate courses while they are still in the Foundation program, so they pass the Foundation program with a numeric GPA. Since the size of this group is small (n=14), we will not address this group and will pay more attention to the large dataset.

5.3 More Data Science Algorithms

As we see from the above discussion, we managed to correctly classify only 73% of the dataset of students enrolled in term 2 in the Foundation Program and not exempted.

In next section we will try other algorithms to explore the possibility of getting better results such as:

- Regression tree
- Random Forest

5.3.1 Regression Tree

The regression tree is formed using CHAID method, with maximum tree depth:3 and the significance level 5%. Since the dataset includes 139 rows, we used 89 (65%) to train the model and the remaining 50 (35%) for testing the model. This division came after several trials of dividing the datasets to reach a reasonable output.

After forming the regression tree, we identified the confusion matrix for the testing dataset, and the results are presented below.

Table 5. 9 Confusion Matrix (Testing Dataset)

			Actual	
Predicted	Fail	Not Fail	Total	% correct
Fail	20	2	22	90.909
Not Fail	10	18	28	64.286
Total	30	20	50	76.000

And the extracted rules as shown in the table below.

Table 5. 10 Tree Rules

Nodes	Status(Pred)	Rules
Node 1	Not Fail	
Node 2	Fail	If High School % <= 85 then Status = Fail in 69.7% of cases
Node 3	Not Fail	If High School % > 85 then Status = Not Fail in 30.3% of cases

As seen from the above table of the extracted rules, if the high school is less or equal 85, then the academic status of students is fail and this is explained for 59.7% of the dataset whereas if the high school is greater than 85% then the status is Not Fail. This conclusion explained for 30% of the dataset.

5.3.2 Random Forest

Random Forest is providing predictive models for classification and regression. It implements binary decision trees. Random forest applied with 70% of the data to train the model and the remaining 30% to test the model. Bagging method used and the number of trees built is 500. The confusion matrix is formed as follows:

Table 5. 11 Confusion Matrix

		Actual		
Predicted	Not Fail	Fail	Total	% correct
Not Fail	56	18	74	75.676
Fail	19	46	65	70.769
Total	75	64	139	73.381

The figure below illustrates the variable importance and seen; the main two variables that contribute to the model are student's performance in Algebra course as well as the high school result.

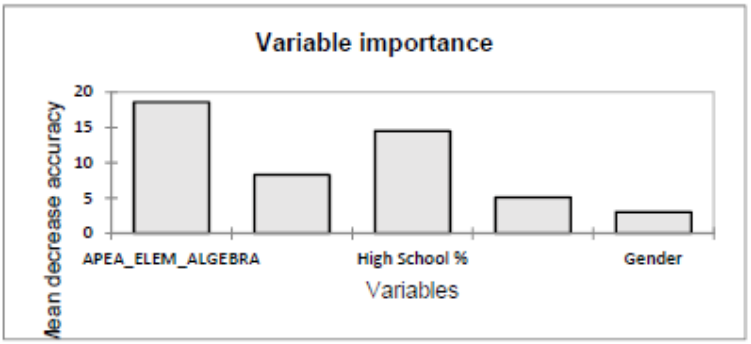


Figure 5. 9 Variable importance

5.4 Summary

The previous discussion addressed the possibility of building predictive model to predict students' academic status at the Foundation level, admitted in term 1 or term 2. For each term the dataset divided into two groups. The first group is related to students are not exempted from

the Foundation program and fully enrolled in the Foundation's courses, and in this case, the logistic regression carried out to predict the binary categorical variable (Fail/ Not Fail). The second group was related to students who were partially exempted and allowed to register undergraduate courses that have letter grade, and hence ended up with a numeric GPA. Since the number of students admitted in both terms, and partially exempted is small number compared with students fully enrolled in the Foundation, we focused on students who are fully enrolled in the Foundation program.

In summary: the table below summarizes the outcomes as follows:

Table 5. 12 Outcomes of students joining the university in term 1 &2

Group	Key factors Admitted in Term 1	Key factors Admitted in Term 2
Foundation level Not exempted from the Foundation courses (registered only Foundation courses (pass/ fail courses with no GPA)	APIC-Integ-Core APEA- Elem-ALGEBRA (88% of the variability of the dependent variable is explained by the explanatory variables).	APEA-Elem-ALGEBRA Age APLA-Arithmetic High School (73.38% of the variability of the dependent variable is explained by the explanatory variables). Age is the most influential.

Although the Foundation Program is offering different levels Mathematics/ English courses to enhance students' skills, the results showed that for students enrolled in term 1, the English course named: APIC-Integ- Core: Integrated Core Elementary, is adding value to the model. This course, according to the University Course Catalogue, is a beginner level integrated skills course aiming to preparing students to study in medium environment. It is listed and defined to be offered 9 lecture hours weekly.

In terms of Mathematics, the results showed that the course named APIEA-Elem-ALGEBRA: Elementary Algebra which is offered 4 lecture hours weekly is affecting students' academic status. This course aiming to enhance students' skills in basic mathematics concepts.

So, only two Foundation courses among all courses taken are affecting the academic status. The situation is different for students admitted in term 2 and fully enrolled in the Foundation Program. The results showed that English courses are not affecting the model. Instead, two Foundation mathematics courses are adding value and affecting the academic status. These courses are:

- APEA- Elem-ALGEBRA
- APEA-Arithmetic.

In addition, high school score has a relationship with the students' performance at the Foundation level.

For the first model related to students admitted in term 1, the independent variables explained 88% of the variability in the dependent variable which is a good result. Other data science algorithms applied to explore the possibility of getting better result for model 2 related to students admitted in term 2. Three data science algorithms were applied: Logistic Regression, regression trees and Random Forest.

Regression trees performed better when compared with the other two methods. Also, it was noticed for students admitted in term 1, only one English / Mathematics courses adding value to the model among other Foundation courses. For students admitted in term 2, the main contributors are Mathematics courses only as well as High school results and all English Foundation courses are contributing to the model.

The next chapter will address the majority and the larger dataset that related to Undergraduate students in order to identify how to predict their GPA, and the main factors affecting the GPA during their academic journey.

Chapter 6 Implementation of Machine Learning Algorithms on Undergraduate students

6.1. Introduction

As mentioned before, the university admit students into two levels. For Science, Engineering, Health, Pharmacy and Medicine Colleges, students need to finish the Foundation program. For other colleges (Arts, Business & Economic, Education, Islamic Studies and Law), the Foundation program is not mandatory, and students affiliated to these colleges are admitted as Undergraduate (UG), where they are eligible to register a UG courses with letter grades. These students will get a numeric GPA by the end of each term. So, the next sections address this group of students and the work focus on exploring the possibility of building predictive models in order to predict students' GPA. The work started with using multiple linear regression, and other data science algorithms are applied and compared as illustrated in coming sections.

Before applying data science techniques, the dataset must be pre-processed including how to handle the missing data if any, transformation of the variables as needed, based on the algorithm that will be used. Since the datasets provided by the University include students in different levels as some students have two term GPA, while others have four term GPA, and others have six term GPA, and in order to compare like with like and treat same group of students equally, there is a need for pre-processing step here which mainly focused on regrouping students as follows:

Group 1: students who have 2_term GPA

Group 2: students who have 4_term GPA

Group 3: students who have 6_ term GPA

For each group, students' historical results used as independent variables to predict the next GPA. For example, for students who have 2_term GPA, the goal is to use any previous results at school and during term 1 in order to predict term_2 GPA

The following chart explains the process as illustrated below.

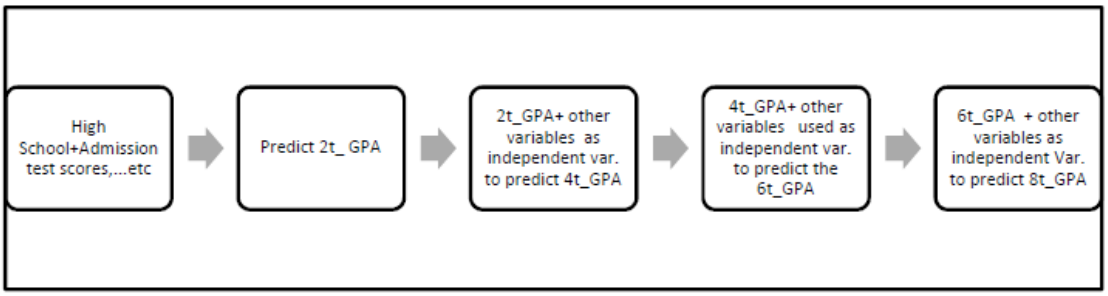


Figure 6. 1 The flowchart of the process of predicting the student's GPA

The above process applied for students admitted in term 1 and term 2 individually.

Before starting to apply the data science algorithms, there is a need for data pre-processing steps, so we handle missing data (if any) as well as transformation of some variables as needed by the algorithm.

6.2. Data pre-processing

6.2.1. Cleaning and integration of datasets

There was a need for restructuring the dataset and grouping of students according to their colleges and disciplines. Also, as part of the data preparation process, the list of students regrouped based on the term GPA and their levels. Splitting the dataset as part of this process in order to treat each group separately based on their levels, admission term, students with 2_term GPA, students with 4_term GPA, and those with 6_term GPA.

6.2.2. Handling missing data

This step is very important before starting the process of building the model. One option to handle missing values is to ignore and remove missing instances. But since in some cases, the missing values might add value to the model. Priya and Sivaraj (2015) conducted a brief review papers that published during the past ten years to identify the methods of handling missing data. Based on this review, it has been shown that Multiple Imputations (MI) is the best and considered by most of the researches as it is more efficient. The process is consisting of three steps. These are: first; imputed m missing values from m complete datasets without any missing values. Then, m datasets are analysed, and the last step is combining the results from m complete datasets. In this process predicted values called "imputes" are substituted (replaced) for missing values.

By performing this process multiple times so all missing values are predicted based on and by using existing values.

6.3. Implementation of data science algorithms on students admitted in Term 1 (Fall)

6.3.1. Group 1: Students with 2_term GPA

The implementation process started with the list of undergraduate students who have two term GPA, and the main goal as mentioned before is to use the historical data and scores as well as the GPA of term 1 to predict the GPA of term 2.

The size of the dataset of students who have 2_term GPA (Group 1) consists of 1411 students. It includes 19 independent variables related to students' historical results in high school,

Foundation Program results, as well as the gender, age, and the GPA of term 1, while the dependent variable is the GPA by the end of term 2.

Three data science algorithms were applied to explore the possibility of predicting students' GPA by the end of term 2, and to identify the main contributors that affect the model.

The process started with the Multiple Linear Regression and the next section address the outcomes of the regression model. Size of the dataset: 1414 students.

Step 1: Test of normality of the dependent variable (Term_2 GPA)

Shapiro- Wilk was applied to test the normality of the dependent variable. The following hypotheses used:

H0: The variable from which the sample was extracted follows a Normal distribution.

H1: The variable from which the sample was extracted does not follow a Normal distribution.

Table 6. 1 Shapiro-Wilk test (Term_2)

W	0.912
p-value (Two-tailed)	< 0.0001
alpha	0.05

As seen from the above table as the computed p-value is lower than the significance level alpha = 0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H1.

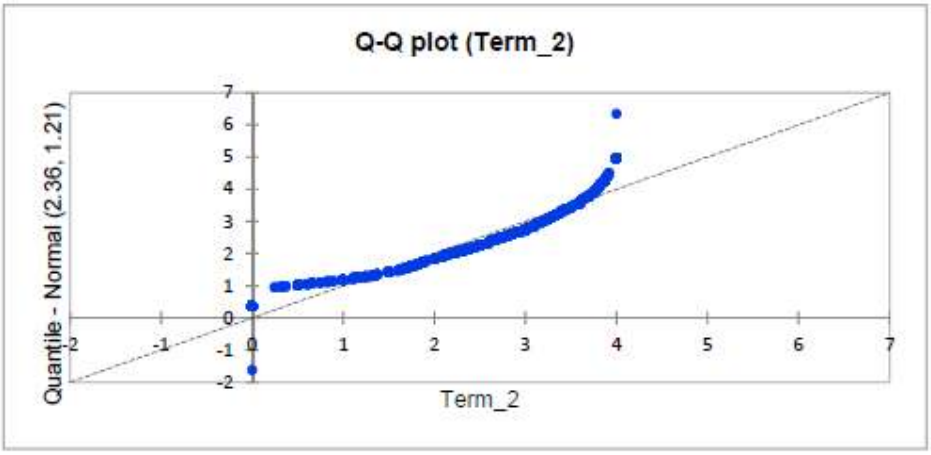
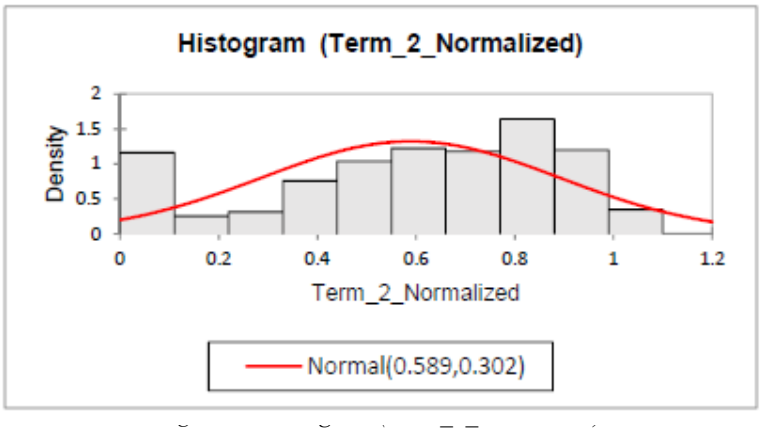


Figure 6. 2 Q-Q plot (Term_2)

Step 2: Normalizing the dependent variable

Since the dependent variable is not normalized, and before starting the regression model, the normalization process was performed. Min-Max Normalization process applied. The dependent variable only rescaled from 0 to 1 to prepare it for the linear regression.

The histogram below shows the dependent variable after performing the normalization process.



Step 3: Correlation matrix in relation to the dependent variable

The correlation matrix was carried out in relation to the dependent variable so that possible independent variables with low correlations can be weeded out early on before considering factor analysis.

Table 6. 2 Correlation matrix (Pearson)

Variables	TOEFL	SAT	IC3	ACT	IELTS	Age	PL_Accuplac	earned_Hour	COLL_Level	Term_2_Nor	Term_1	igh School %	LA_ Arithme	_ELEM_ALGS	_Reading_	PLG_listenin	LU_Lang_U	Sentence_M	Writing_Wo	IC_Integ_Co
TOEFL	1	0.000	0.110	-0.054	-0.015	-0.006	0.118	0.046	0.000	0.008	-0.014	-0.002	0.001	0.002	-0.002	0.002	0.000	-0.001	0.000	0.000
SAT	0.000	1	0.044	-0.016	0.073	0.040	0.041	0.022	0.000	0.027	0.026	0.050	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
IC3	0.110	0.044	1	0.087	0.005	0.019	0.292	0.079	0.000	0.061	0.067	-0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ACT	-0.054	-0.016	0.087	1	0.183	-0.056	-0.058	-0.061	0.004	0.026	0.017	0.080	-0.023	-0.017	0.014	0.003	0.014	0.010	0.013	0.012
IELTS	-0.015	0.073	0.005	0.183	1	-0.019	-0.066	0.073	0.072	0.092	0.081	0.106	-0.004	0.029	-0.034	-0.020	-0.027	-0.026	-0.027	-0.029
Age	-0.006	0.040	0.019	-0.056	-0.019	1	0.019	-0.048	-0.017	-0.223	-0.195	-0.205	-0.044	-0.079	-0.061	-0.041	-0.049	-0.047	-0.042	-0.047
APL_Accu	0.118	0.041	0.292	-0.058	-0.066	0.019	1	0.093	0.016	0.062	0.068	0.027	-0.008	0.042	0.084	0.060	0.100	0.081	0.010	0.011
Earned_H	0.046	0.022	0.079	-0.061	0.073	-0.048	0.093	1	0.075	0.630	0.627	0.261	0.062	0.188	0.060	0.059	0.061	0.062	0.061	0.064
APCL_CC	0.000	0.000	0.000	0.004	0.072	-0.017	0.016	0.075	1	0.100	0.142	0.140	-0.002	0.159	0.116	0.083	0.126	0.101	0.119	0.119
Term_2_N	0.008	0.027	0.061	0.026	0.092	-0.223	0.062	0.630	0.100	1	0.681	0.451	0.084	0.221	0.129	0.104	0.106	0.096	0.095	0.108
Term_1	-0.014	0.026	0.067	0.017	0.081	-0.195	0.068	0.627	0.142	0.681	1	0.434	0.095	0.225	0.141	0.097	0.142	0.140	0.130	0.128
High Scho	-0.002	0.050	-0.020	0.080	0.106	-0.205	0.027	0.261	0.140	0.451	0.434	1	0.094	0.316	0.195	0.135	0.175	0.164	0.168	0.176
APLA_ Ari	0.001	0.000	0.000	-0.023	-0.004	-0.044	-0.008	0.062	-0.002	0.084	0.095	0.094	1	0.217	0.279	0.232	0.234	0.278	0.276	0.286
APEA_EL	0.002	0.000	0.000	-0.017	0.029	-0.079	0.042	0.188	0.159	0.221	0.225	0.316	0.217	1	0.338	0.253	0.329	0.298	0.321	0.329
APRS_Re	-0.002	0.000	0.000	0.014	-0.034	-0.061	0.084	0.060	0.116	0.129	0.141	0.195	0.279	0.338	1	0.777	0.789	0.817	0.821	0.906
APLG_list	0.002	0.000	0.000	0.003	-0.020	-0.041	0.060	0.059	0.083	0.104	0.097	0.135	0.232	0.253	0.777	1	0.815	0.827	0.850	0.901
APLU_Lar	0.000	0.000	0.000	0.014	-0.027	-0.049	0.100	0.061	0.126	0.106	0.142	0.175	0.234	0.329	0.789	0.815	1	0.841	0.938	0.910
APSM_Se	-0.001	0.000	0.000	0.010	-0.026	-0.047	0.081	0.062	0.101	0.096	0.140	0.164	0.278	0.298	0.817	0.827	0.841	1	0.944	0.926
APWS_W	0.000	0.000	0.000	0.013	-0.027	-0.042	0.010	0.061	0.119	0.095	0.130	0.168	0.276	0.321	0.821	0.850	0.938	0.944	1	0.976
APIC_Inte	0.000	0.000	0.000	0.012	-0.029	-0.047	0.011	0.064	0.119	0.108	0.128	0.176	0.286	0.329	0.906	0.901	0.910	0.926	0.976	1

As seen from the above matrix, the highly correlated variables in relation to the dependent variable" Term_2 GPA" are:

- Term_1 GPA (0.681)
- Earned hours (0.63)
- High school % (0.451)

Another indicator that some variables need to be removed is the KMO test which was performed as another indicator. We looked at the individual variables below 0.5 as candidates for removing. These variables compared with the correlation matrix.

The eigenvalues below show the variation for each factor. Based on Kaiser (1960), all eigenvalues greater than 1 considered.

After the varimax rotation, high loading scores were considered as these values indicate the variable representatives of the factor.

Linear Regression

From each group, and for those highly correlated variables, a representative variable was selected to represent that group and to be included in the model instead of dropping all variables

in the model. In order to select a representative from each group, three outputs were compared. These are:

- An individual score of each variable in the Kaiser-Meyer-Olkin (KMO) measure,
- The correlation matrix in relation with the dependent variable, and
- Factor loading scores for each variable under each factor, and representative variables were selected. The selection made based on the variable that has highest correlation with the dependent variable.

After reviewing the above three outputs, the results show that some variables can be weeded out early and not included in the models.

Following the same process for other variables and after comparing the three outputs matrices, the following variables were extracted as follows:

- Earned hours
- Term_1 GPA
- High School %

Multiple Linear Regression

Stepwise linear regression was carried out to predict the GPA of term_2 with Confidence interval 95%. The equation of the built model (Term_2 GPA) is formed as follows:

Term_2 GPA = -0.527156452132665+7.47917221612249E-03*High School %+1

Using the Stepwise variables selection method, 3 variables have been retained in the model. Given the R^2, 56% of the variability of the dependent variable Term_2 GPA is explained by the 3 explanatory variables.

Regression Tree

The Regression tree has been built using "rpart". The dataset was divided as 70% for training the model and 30% for testing. The rules were generated. Further details are illustrated in the appendix. The tree is illustrated below as seen.

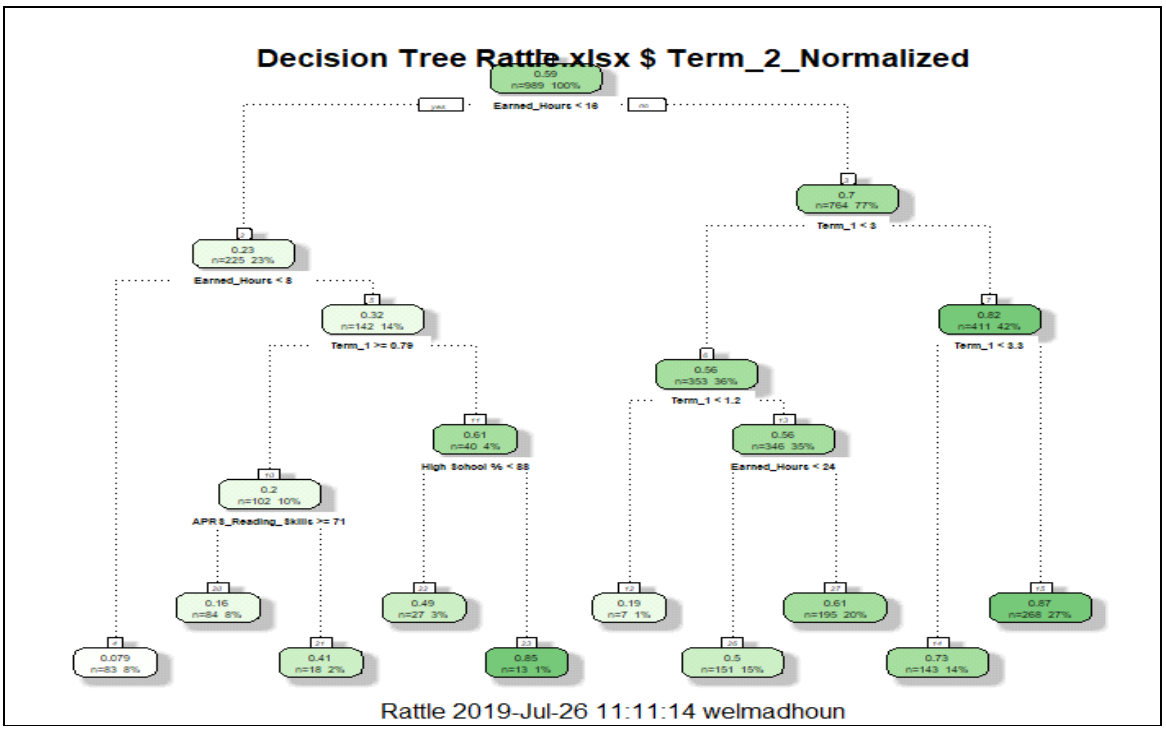


Figure 6. 4 Regression Tree- students admitted in term 1 &2

Random Forest

Random Forest is building multiple decision trees from different samples from the dataset. During this process random subsets of each variable are used for splitting the data at each node. The method was implemented out and instead of building one tree, it works on building several trees and combine their predictions. A summary statistic (Training / Quantitative) presented in the appendix. The graph below shows the most important variables that affect the model as follows:

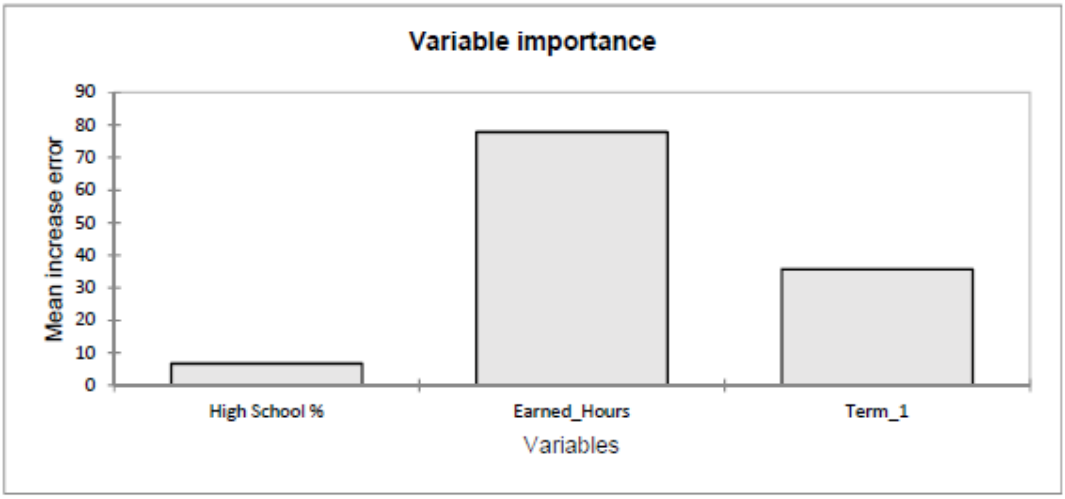


Figure 6. 5 Variable importance

The most important variables that affect the model in order (highest to lowest) are:

1) Earned Hours

2) Term_1 GPA

3) High School %.

6.3.2. Group 2: Students with 4_term GPA

The size of the dataset of this group is 1375 students. Their age is between 17 and 47 years. The dataset for this group includes students' historical scores and results from the Foundation Program as well as their GPA in the previous terms (Term_1, term_2, term_3), and the goal is to explore the possibility of predicting the fourth term GPA (term_4 GPA) by using the mentioned historical results. Since the dataset includes 19 variables, some of them might not needed to build the model, and should be wedded up, so the same approach was followed as shown above for dataset of students who have 2 term_GPA. The process started with testing the normality of the dependent variable as seen below.

Normality test of the dependent variable

Before starting the regression model, we tested the normality of the dependent variable which is term 4_GPA. Shapiro-Wilk test used with 5% significance level. The null hypothesis H0 is the dependent variable follows a normal distribution, and the alternative hypothesis H1 is the dependent variable does not follow a Normal distribution. The test carried out and the result are illustrated in the appendix. As the computed p-value is lower than the significance level alpha 0.05, one should reject the null hypothesis H0, and accept the alternative hypothesis H1.

Max-Min method used to normalize the dependent variable and the normalized variable was used in the model built. After normalizing the dependent variable, the shape of the histogram is presented below as seen.

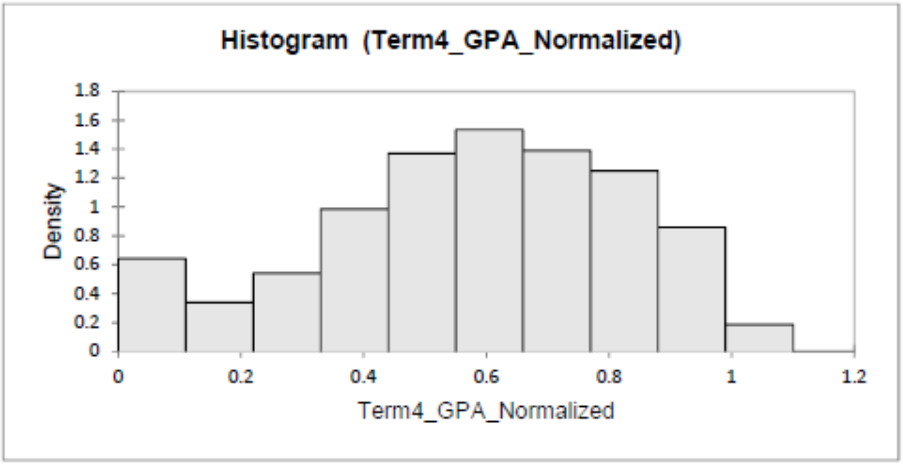


Figure 6. 6 Histogram after normalizing the dependent variable

After normalizing the depending variable we started the process with testing the correlation between the depending variable and other variables from one side and between dependent variables from the other side. The idea behind this is if the correlation between independent variables is high, then there is no need to include all variables in the model, and we will shift the direction to factor Analysis to extract variables that need to be included in the model.

Correlation test

Pearson correlation test performed to explore the correlation between the dependent variable and independent variables. The table below illustrates the correlation matrix obtained.

Table 6. 3 Correlation Matrix (Pearson)

Variables	APCI_COLL_I	IELTS	APLG_listenin	APRS_Reading	APIC_Integ_C	APWS_Writing	APSM_Senten	APLU_Lang_U	Age	High School %	Term2_GPA	Term3_GPA	Earned_Hours	Term4_GPA_N	Term1_GPA	APEA_ELEM_/	APLA_Arithme
APCI_COLL_Level_M	1	0.008	-0.018	-0.003	-0.001	0.002	-0.023	-0.013	0.013	-0.007	0.032	0.044	0.040	0.024	0.035	-0.128	-0.020
IELTS	0.008	1	-0.052	-0.048	-0.057	-0.038	-0.039	0.006	-0.039	0.058	0.079	0.060	0.099	0.060	0.071	0.007	-0.006
APLG_listening	-0.018	-0.052	1	0.416	0.604	0.488	0.492	0.450	-0.036	0.136	0.022	0.030	0.047	0.039	0.073	0.212	0.104
APRS_Reading_Skills	-0.003	-0.048	0.416	1	0.660	0.562	0.539	0.546	-0.024	0.172	0.055	0.053	0.067	0.059	0.107	0.284	0.107
APIC_Integ_Core	-0.001	-0.057	0.604	0.660	1	0.932	0.505	0.492	-0.024	0.155	0.050	0.054	0.075	0.102	0.120	0.252	0.013
APWS_Writing_Words	0.002	-0.038	0.488	0.562	0.932	1	0.608	0.610	-0.039	0.153	0.052	0.054	0.083	0.115	0.126	0.246	0.023
APSM_Sentence_Mec	-0.023	-0.039	0.492	0.539	0.505	0.608	1	0.546	-0.045	0.146	0.030	0.028	0.060	0.033	0.072	0.230	0.142
APLU_Lang_Use	-0.013	0.006	0.450	0.546	0.492	0.610	0.546	1	-0.080	0.176	0.023	0.030	0.059	0.063	0.099	0.234	0.160
Age	0.013	-0.039	-0.036	-0.024	-0.024	-0.039	-0.045	-0.080	1	-0.281	-0.098	-0.052	-0.068	-0.057	0.009	-0.065	-0.059
High School %	-0.007	0.058	0.136	0.172	0.155	0.153	0.146	0.176	-0.281	1	0.435	0.450	0.453	0.471	0.471	0.290	0.105
Term2_GPA	0.032	0.079	0.022	0.055	0.050	0.052	0.030	0.023	-0.098	0.435	1	0.634	0.705	0.593	0.528	0.165	0.044
Term3_GPA	0.044	0.060	0.030	0.053	0.054	0.054	0.028	0.030	-0.052	0.450	0.634	1	0.724	0.643	0.597	0.175	0.031
Earned_Hours	0.040	0.099	0.047	0.067	0.075	0.083	0.060	0.059	-0.068	0.453	0.705	0.724	1	0.718	0.703	0.217	0.060
Term4_GPA_Normaliz	0.024	0.060	0.039	0.059	0.102	0.115	0.033	0.063	-0.057	0.471	0.593	0.643	0.718	1	0.623	0.233	-0.015
Term1_GPA	0.035	0.071	0.073	0.107	0.120	0.126	0.072	0.099	0.009	0.471	0.528	0.597	0.703	0.623	1	0.236	0.098
APEA_ELEM_ALGEBI	-0.128	0.007	0.212	0.284	0.252	0.246	0.230	0.234	-0.065	0.290	0.165	0.175	0.217	0.233	0.236	1	0.256
APLA_Arithmetic	-0.020	-0.006	0.104	0.107	0.013	0.023	0.142	0.160	-0.059	0.105	0.044	0.031	0.060	-0.015	0.098	0.256	1

As seen from the above matrix, there are some correlations between independent variables themselves and to avoid including all of them in the model, the section below discuss and present the factor analysis process that have been performed to extract the main variables that need to be included in the model.

Factor Analysis was performed and included all independent variables (the target variable was not part of this process). Varimax rotation used to identify the best positions to variables and their associations to factors.

As we did with the previous group of students with 2_term GPA, we looked into the correlation matrix, and compared an individual score of KMO. Any variable scored below 0.5 will be excluded. We also compared the factor loading scores and variables association with factors. We consider those scored 0.7 and above.

After comparing the correlation matrix, KMO scores and factor loading scores as well as variables associations with factors, a representative variable from each factor was selected instead of including all variables in the model. After several trials we ended up with the following variables that contribute to the model as follows:

- High School %
- Earned Hours
- Term _1 GPA
- Term_2 GPA
- Term _3 GPA

Regression Model

After several experiments and by using stepwise model to add and remove variables, the correlation matrix is obtained and presented below:

Table 6. 4 Correlation Matrix

	Term1_ GPA	Term2_ GPA	Term3_ GPA	High School %	Earned_ Hours	Term4_ GPA_
Term1_GPA	1	0.505	0.591	0.469	0.694	0.612
Term2_GPA	0.505	1	0.640	0.433	0.701	0.591
Term3_GPA	0.591	0.640	1	0.448	0.719	0.649
High School %	0.469	0.433	0.448	1	0.454	0.474
Earned_Hours	0.694	0.701	0.719	0.454	1	0.722
Term4_GPA_Normalized	0.612	0.591	0.649	0.474	0.722	1

And the Type I Sum of Squares analysis (Term4_GPA)

Table 6. 5 Type I Sum of Squares analysis (Term_4 GPA)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Term1_GPA	1	29.603	29.603	1057.078	< 0.0001
Term2_GPA	1	8.454	8.454	301.888	< 0.0001
Term3_GPA	1	4.008	4.008	143.124	< 0.0001
High School %	1	0.831	0.831	29.689	< 0.0001
Earned_Hours	1	3.400	3.400	121.390	< 0.0001

Type III Sum of Squares analysis (Term4_GPA)

Table 6. 6 Type III Sum of Squares analysis

Source	DF	Sum of squares	Mean squares	F	Pr > F
Term1_GPA	1	0.774	0.774	27.649	< 0.0001
Term2_GPA	1	0.271	0.271	9.683	0.002
Term3_GPA	1	1.205	1.205	43.015	< 0.0001
High School %	1	0.751	0.751	26.803	< 0.0001
Earned_Hours	1	3.400	3.400	121.390	< 0.0001

According to the Type III sum of squares, the following variables bring significant information to explain the variability of the dependent variable Term4_GPA are as follows:

- Earned_Hours
- Term_3 GPA
- Term_1 GPA
- High School %

Among the explanatory variables, based on the Type III sum of squares, variable Earned Hours is the most influential.

Equation of the model (Term4_GPA)

Term4_GPA_Normalized = -0.33220693769111+3.54388303055351E-02*Term1_GPA+1.90424467561203E-02*Term2_GPA+4.58169508110564E-02*Term3_GPA+4.72504218571661E-03*High School %+6.54912092324064E-03*Earned_Hours

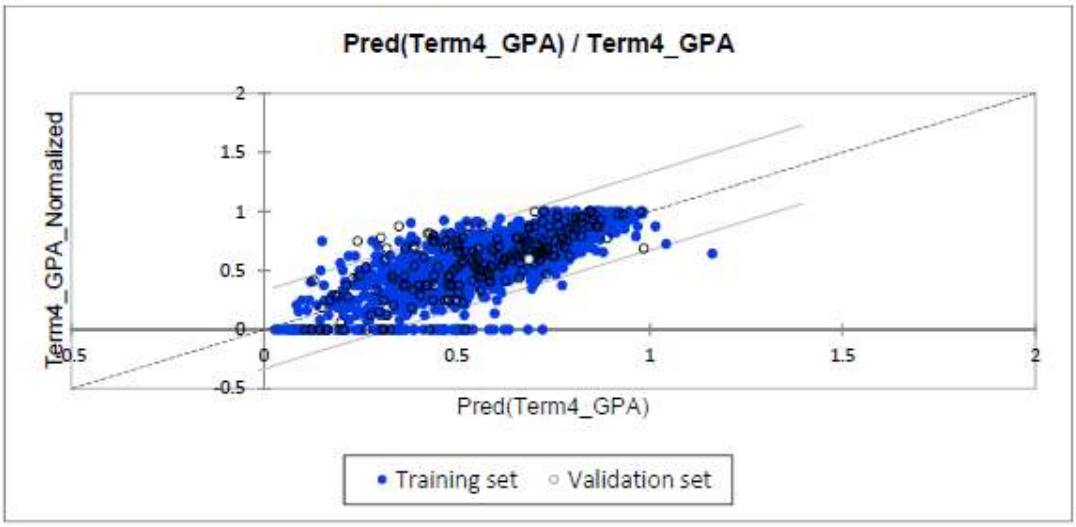


Figure 6. 7 Predicted Term_4 GPA

Based on R-Square, 59% of the variability in the dependent variable explained by the three exploratory variables, and this is the maximum percent we reached after several trials. Another Machine Learning Method applied to explore the possibility to improve the model.

Regression Tree

Now, as before, the regression Tree formed. 70% of the data used for training the model and the remaining for testing. The tree structure is presented below.

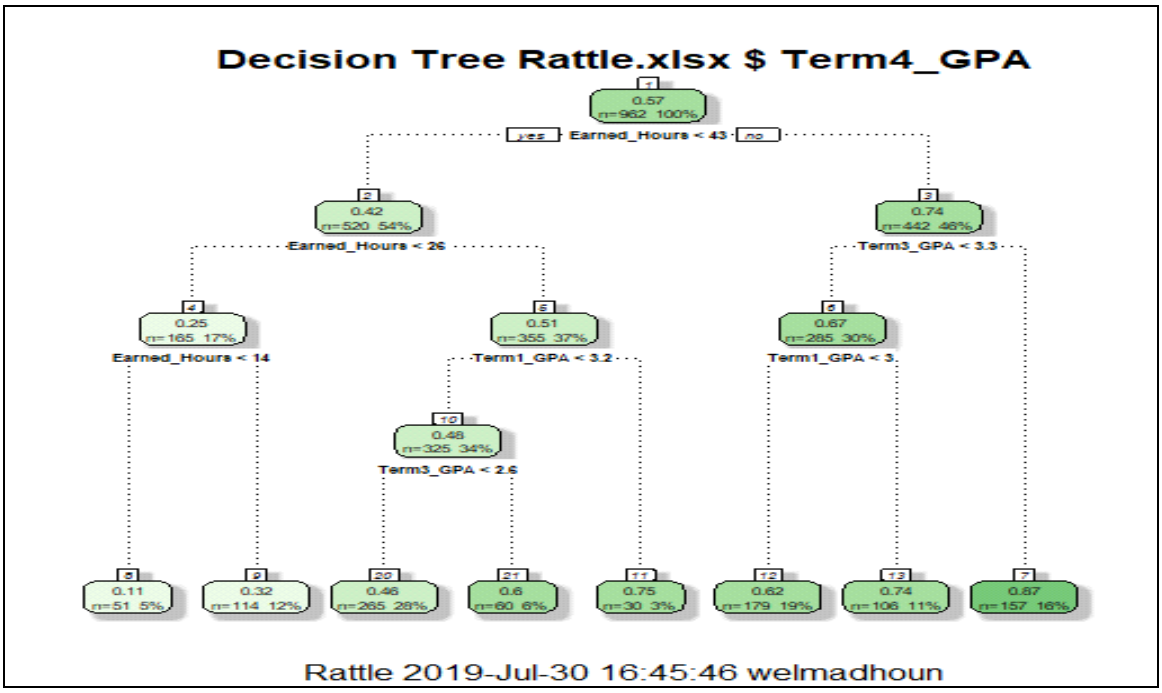


Figure 6. 8 Tree structure

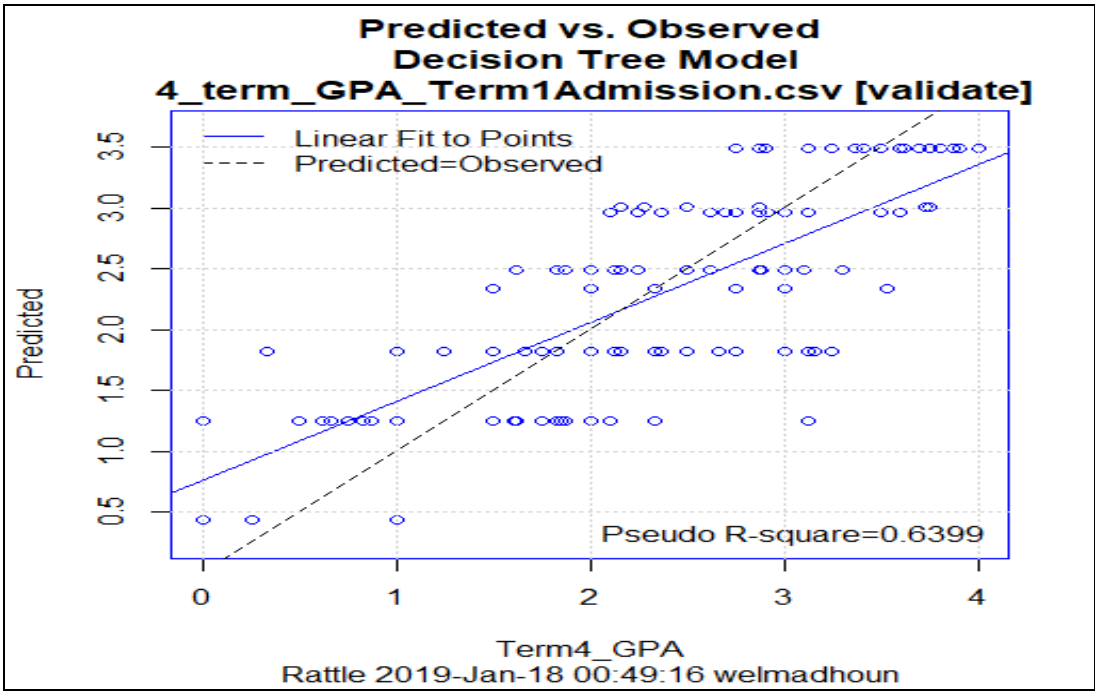


Figure 6. 9 Predicted vs. Observed values

As appears in the figure, the Pseudo R-Square is 0.64 which is square of the correlation between the predicted and observed values. The closer to 1, the better the built model.

Random Forest

Now as done in the previous group of students, Random Forest method performed to extract the variable importance. Further details about the model in the appendix. The main contributors to the model are:

- Earned hours
- Term_3 GPA
- Term_2 GPA
- High School %

The next section will address the implementation of the same three models on the list of students who have 6_ term GPA.

6.3.3. Group 3: Students with 6_term GPA

The third group is for students who spent six terms at the university. The dataset includes 103 students. It includes students' scores in several courses as well the term GPA for the first five terms (term_1 to term_5). There are 19 variables. The author explored the possibility of predicting term_6 GPA by using the independent variables and following the same approach followed for the previous two groups of students.

As above the process started with testing the correlations among the dependent variables and independent variables as well as between independent variables themselves to explore the possibility of performing Factor Analysis and extract certain variables that contribute to the model.

Further details about the Factor Analysis process, the varimax rotation and the extracted variables are illustrated in Appendix.

A representative from each factor extracted. Although the author looked into all variables scored 0.7 and above but also considered the representative the one that has highest correlation with the dependent variable which is Term_5 GPA.

Linear Regression

After several trials and by using step wise regression method, we ended up with one variable that significantly affect the dependent variable which is term_5 GPA. It is highly correlated with the dependent variable (Term_ 6 GPA) and the correlation is 0.99.

70% of the data used for training the model and 30% for testing. After several trials, it concluded that the only variable that significantly affects the model is the GPA of term 5. Further details are illustrated in the appendix.

By Using the Stepwise variable selection method, one variable has been retained in the model. **The equation** of the models is given as follows:

Term_6_GPA = -0.199484936213633+0.298142377861267*Term_5

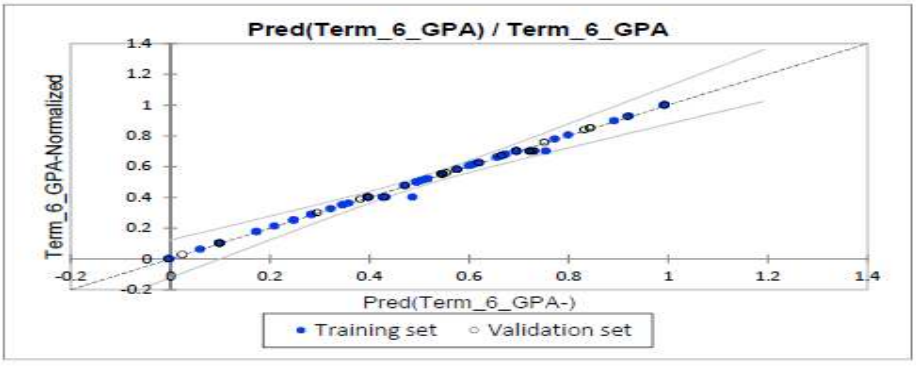


Figure 6. 10 Predicted Term_6 GPA

Full details about the above model are presented in Appendix.

Regression Trees

The regression tree formed using 'rpart' . 70% of the data used to build the model and the remaining 30% used to test the model. The rules generated and the main variable used to build the tree is the GPA of term_5, and the structure of the tree is illustrated below:

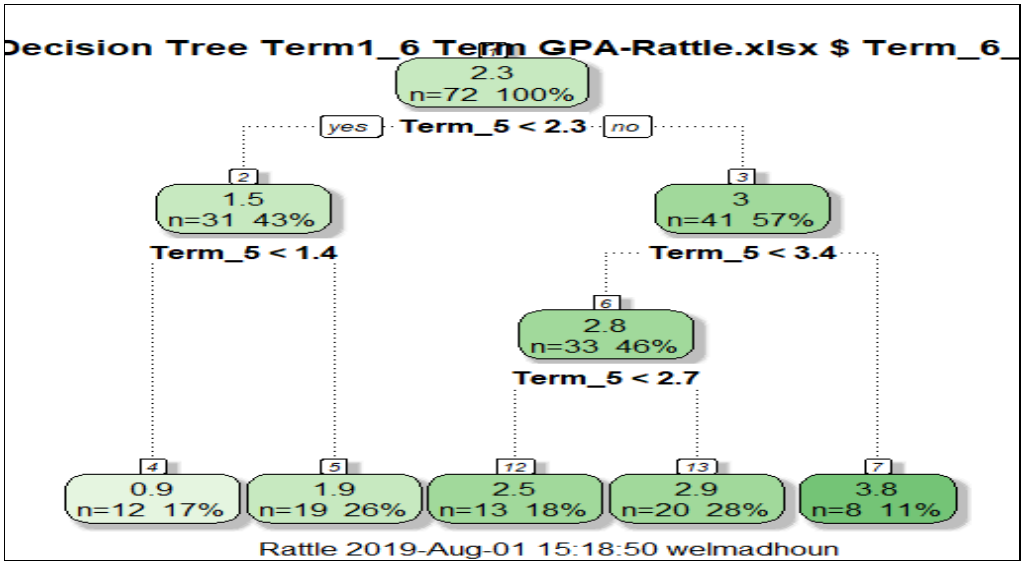


Figure 6. 11 Regression Tree (Term_6 GPA)

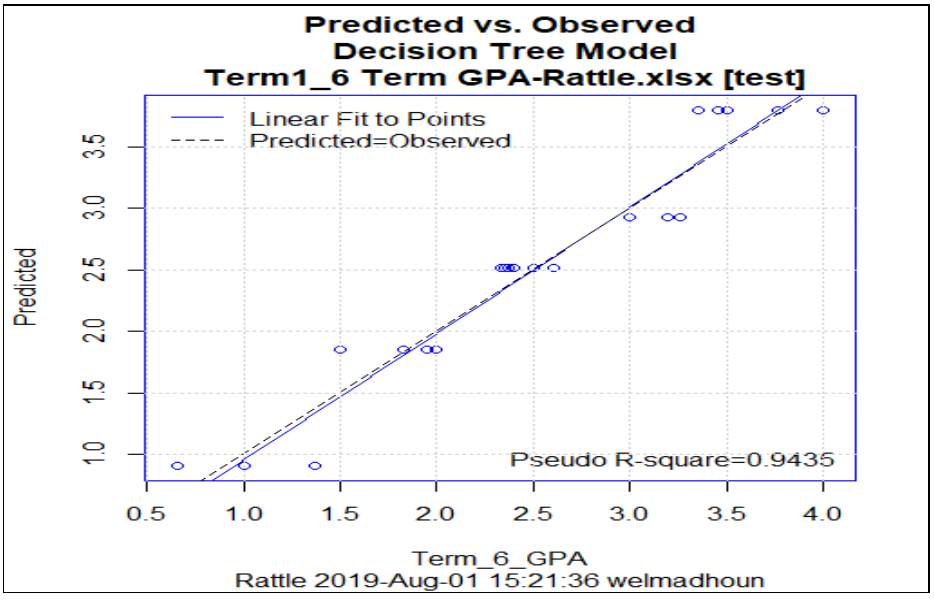


Figure 6. 12 Predicted vs. observed values- students with 6_term GPA and admitted in term 1

The predicted vs. observed values shows the Pseudo R-Square as well which is 0.94 which is very high score.

Random Forest

As previous sections, the author carried out Random Forest model. The number of trees used to build is 500 trees. Also, as previous models, 70% of the data used to train the model. The output is illustrated in the appendix.

Summary of the Random Forest Model

==========

randomForest(formula = Term_6_GPA ~ .,

data = crs$dataset[crs$train, c(crs$input, crs$target)],

ntree = 500, mtry = 1, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 1

Mean of squared residuals: 0.05244518

% Var explained: 93.25

The error rate and the variable importance are also presented below as seen.

Variable Importance

Importance

Term_5 0.6598966

Earned_Hours 0.2849408

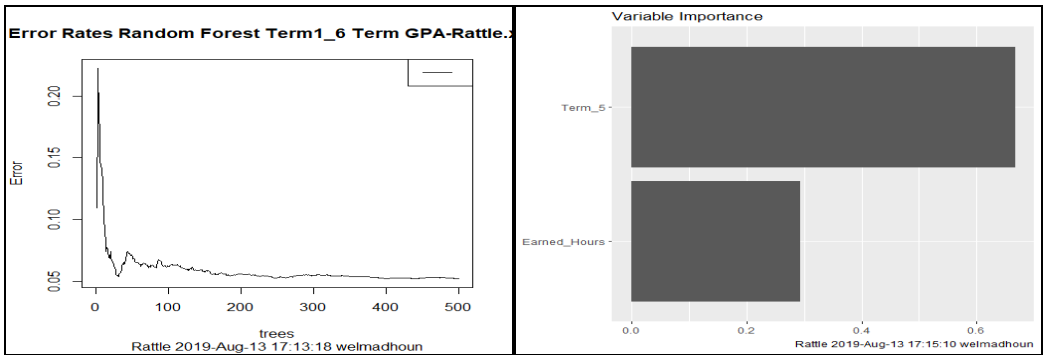


Figure 6. 13 Error rate and variable importance- students with 6_term GPA and admitted in term 1

The graph above shows that the more the number of trees used, the lower the error rate.

6.3.4. Summary

In the above discussion, and for the three groups of students, three data science algorithms implemented for each group of students (students who have 2_term GPA, Students who have 4_term GPA and students who have 6_term GPA).

The above discussion showed that among the three Machine Learning methods applied, the most important factors that contribute to the model are mainly, students' previous term GPA. For students who have 2_term GPA, it was concluded that the previous (Term_1 GPA) as well as earned credit hours, and the high school % are the main contributors to the model.

The same above algorithms were repeated for Group 2 (students who have 4_term GPA) and Group 3 (students who have 6_term GPA) in order to identify the main factors that affect students' performance in these levels. The results showed that for students who have 4_term GPA, the main variables that affect the GPA are:

- Earned hours
- Term_3 GPA
- Term_2 GPA
- High School %

While for those who have 6_term GPA, the GPA of the previous term (Term_5 GPA) is most important factor. It was a surprise that none of the Foundation program courses that usually students spend one full academic year to finish adding any value to the GPA. Further investigation and discussion about this important finding will be addressed later in coming sections.

6.4 Implementation of Machine Learning algorithms on students admitted in Term 2 (spring)

In this section we explore the possibility of building predictive models to predict the student's GPA for each group as indicated in previous section. Furthermore, we will try to identify the main factors that affect students' performance either those who have 2_term GPA, 4_term GPA or 6_term GPA.

As mentioned before, the university has a certain capacity for each academic year distributed between the two terms. For those who did not accepted in term 1 due to the enrolment capacity, they usually are advised to apply for term 2 intake.

In order to compare same group of students, and as we did with the dataset of students' joint the university in term 1, the list of students who admitted in term 2 was divided into three groups as we did for those who were admitted in term 1. These groups are:

Group 1: students who have 2_ term GPA (the size of this group is 326 students)

Group 2: students who have 4_term GPA (the size of this group is 261)

Group 3: students who have 4_term GPA (the size is 169).

The above data science algorithms were repeated for each group in order to identify which algorithm is performing better, and what are the main variables that affect the student's GPA for each group.

6.4.1. Group 1: Students with 2_term GPA

When analysing the results of machine learning algorithms related to the dataset of students admitted in term1 (Fall), and term 2 (Spring), we can see that one key and common conclusion for both terms is that the students' performance (despite the number of terms the student spent in the university) is not affected by their performance in the Foundation Program that designed mainly to prepare students to their majors. Although the program was designed to give students affiliated to Engineering, Science, Health, Medicine, and Pharmacy colleges the competency in Mathematics and English (oral/ written) skills to the academic standards at the concerned university. Further elaboration about this key observation will be addressed in detail in coming section as well as further explorations. Since the dataset includes 18 variables, how to identify the variables need to be included in the model, the same approach used in previous section followed which starts with correlation test, and for the highly correlated variables we moved to factor analysis including all independent variables to identify the factors and variables association with factors. A representative variable was selected, and key variables were included in the model instead of dropping out all variables in the model.

Pearson correlation test performed, and the matrix below illustrates the correlation among independent variables as follows:

Table 6. 7 Correlation Matrix (Pearson)- Independent variables

Variables	APCL_COL	IELTS	Term2_GP	Earned_Ho	Term1_GP	High Schoo	APEA_ELE	APRS_Rea	APIC_Integ	APWS_Wri	APLU_Lang	APSM_Sen	APLG_liste	APLA_ Arith	APL_Accup	Age
APCL_COLL_Level_MATH	1	0.043	0.063	-0.033	0.044	0.026	0.098	-0.033	-0.049	-0.016	-0.020	-0.068	-0.110	-0.035	-0.017	0.005
IELTS	0.043	1	0.013	-0.010	0.114	0.082	0.128	0.070	0.079	0.089	0.078	0.045	0.054	-0.027	0.020	0.016
Term2_GPA	0.063	0.013	1	0.500	0.496	0.179	0.236	0.049	0.061	0.073	0.036	0.045	0.080	0.025	-0.032	-0.046
Earned_Hours	-0.033	-0.010	0.500	1	0.544	0.177	0.140	0.059	0.124	0.110	0.100	0.116	0.099	0.002	0.087	-0.062
Term1_GPA	0.044	0.114	0.496	0.544	1	0.360	0.227	0.019	0.079	0.077	0.051	0.067	0.085	0.015	-0.023	-0.151
High School %	0.026	0.082	0.179	0.177	0.360	1	0.316	0.231	0.243	0.248	0.183	0.262	0.142	0.140	0.015	-0.228
APEA_ELEM_ALGEBRA	0.098	0.128	0.236	0.140	0.227	0.316	1	0.429	0.420	0.406	0.375	0.303	0.276	0.231	0.044	-0.064
APRS_Reading_Skills	-0.033	0.070	0.049	0.059	0.019	0.231	0.429	1	0.770	0.645	0.585	0.596	0.480	0.264	0.051	-0.038
APIC_Integ_Core	-0.049	0.079	0.061	0.124	0.079	0.243	0.420	0.770	1	0.908	0.759	0.719	0.664	0.297	0.037	-0.050
APWS_Writing_Workshop	-0.016	0.089	0.073	0.110	0.077	0.248	0.406	0.645	0.908	1	0.821	0.781	0.587	0.262	0.022	-0.062
APLU_Lang_Use	-0.020	0.078	0.036	0.100	0.051	0.183	0.375	0.585	0.759	0.821	1	0.601	0.590	0.243	0.069	-0.088
APSM_Sentence_Meaning	-0.068	0.045	0.045	0.116	0.067	0.262	0.303	0.596	0.719	0.781	0.601	1	0.573	0.167	0.061	-0.113
APLG_listening	-0.110	0.054	0.080	0.099	0.085	0.142	0.276	0.480	0.664	0.587	0.590	0.573	1	0.188	0.083	-0.084
APLA_ Arithmetic	-0.035	-0.027	0.025	0.002	0.015	0.140	0.231	0.264	0.297	0.262	0.243	0.167	0.188	1	0.072	0.064
APL_Accuplacer	-0.017	0.020	-0.032	0.087	-0.023	0.015	0.044	0.051	0.037	0.022	0.069	0.061	0.083	0.072	1	-0.088
Age	0.005	0.016	-0.046	-0.062	-0.151	-0.228	-0.064	-0.038	-0.050	-0.062	-0.088	-0.113	-0.084	0.064	-0.088	1

As seen from the above correlation matrix, there is a high correction between some independent variables as highlighted in the above matrix. Based on this and to identify the factors and a

representative variable associated with each factor, factor analysis process repeated, and ended up with only two independent variables that affected the GPA of term 2.

The variable representative that selected those which are highly correlated with the dependent is:

- Term_1 GPA
- Earned hours.

Linear Regression

Stepwise regression model carried out with 70% of the data used to train the model and 30% for testing, and after several trials, we ended up with the above two variables that affect the model.

The table below illustrates the Type III Sum of Squares analysis (Term2_GPA):

Table 6. 8 Type III Sum of Squares analysis (Term _2 GPA)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	1	42.641	42.641	31.297	< 0.0001
Term1_GPA	1	37.462	37.462	27.495	< 0.0001

As seen from Type III sum of squares analysis, the variable that affected the model is the number of earned credit hours and then the previous term GPA (term_1 GPA).

Further details about the model are presented in Appendix.

Regression Trees

As above the regression trees was performed as well, and rules were extracted. The generated rules are illustrated in the appendix. 70% of the data used for training the model and the remaining 30% for testing. The tree structure is presented below.

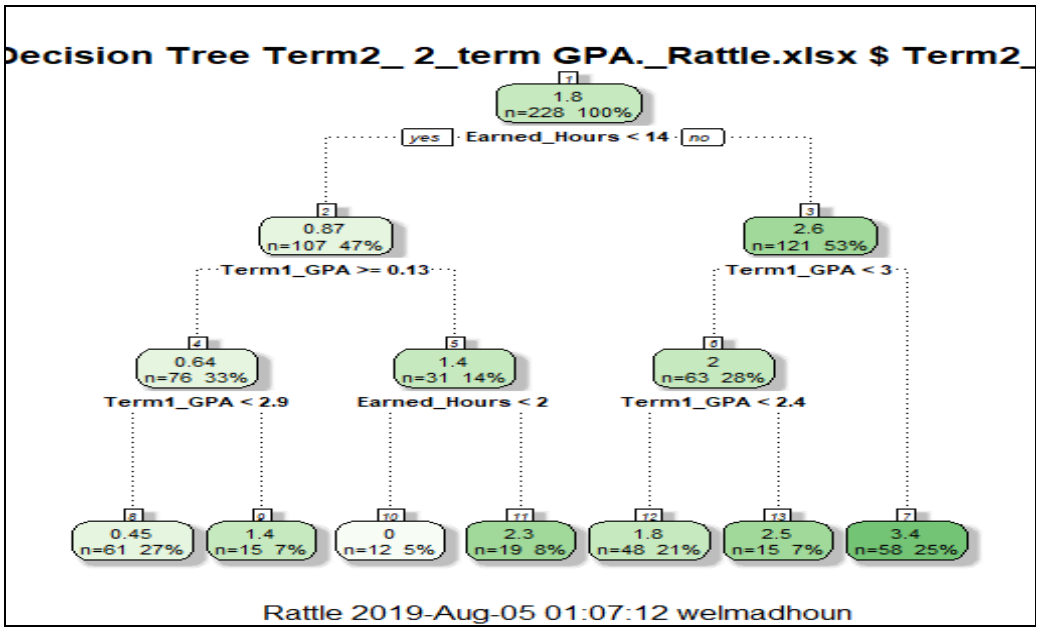


Figure 6. 14 Tree Structure- students who have 2_term GPA (Admitted in Term 2)

Furthermore, to evaluate the model, we checked the predictive vs. observed lines. As we see below, and after several trials, the Pseudo R-Square is 6.63.

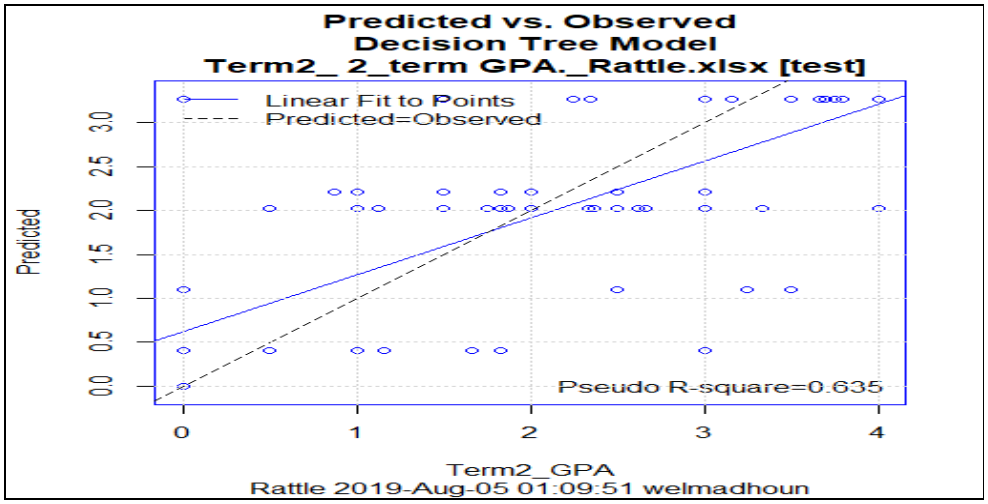


Figure 6. 15 Predicted vs observed values- students admitted in term 2 and have 2_term GPA

Random Forest

Random Forest was carried out as well, and 500 trees used. As above 70% of the data were used to train the model and the remaining 30% to test the model. The outputs are illustrated in the appendix and the variable importance are presented below as follows:

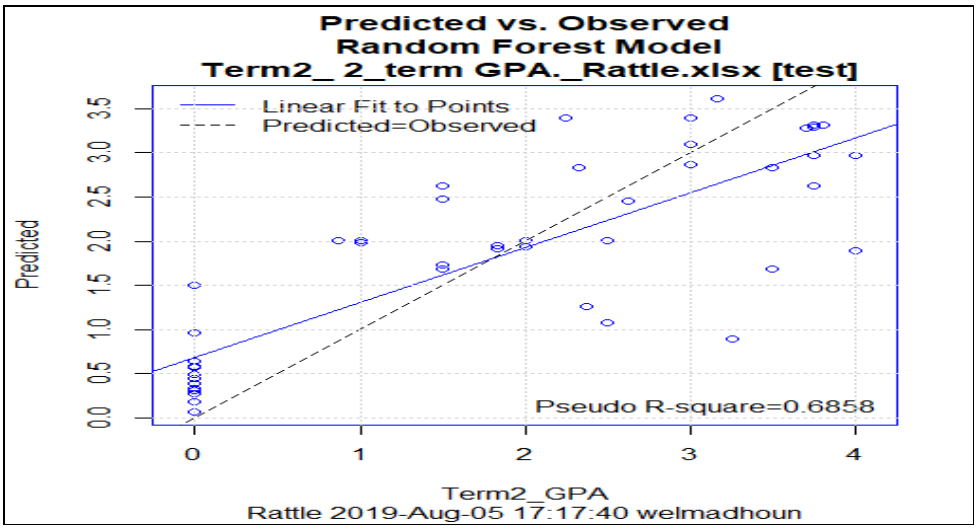


Figure 6. 16 Predicted vs observed values- Students admitted in term 2 and have 2_term GPA, and the second graph shows variable importance

As seen from the above results, the most important variable is the number of earned hours then the GPA of term_1. This result is consistent with the previous model's conclusion.

Furthermore, according to the Pseudo R-Square, the performance of the regression tress (0.63) and random forest model (0.68) is better compared with the linear regression model (0.48).

The next section will present the outputs of implementing the same methods on dataset of students who have 4_term GPA and admitted in term 2.

6.4.2. Group 2: Students with 4_term GPA

The dataset includes 261 students and since there are 19 variables including students' scores in different level courses, the same approach as above followed. Pearson correlation test performed to identify any high correlation between independent variables to select a representative instead of including all the 19 variables in the model.

The process started with exploring the correlation between the dependent variable (Term_4 GPA) and the 19 independent variables. After performing Pearson correlation test the correlation matrix is obtained as indicated below.

Table 6. 9 Correlation matrix (Pearson)- Students with 4_ term GPA and admitted in term 2

Variables	IELTS	APCL_CC	Age	High School %	Term 1_GPA	Term 3_G	Earned_Hours	Term 4_GPA_	Term 2_GPA	APEA_EL	APRS_Re	APSM_Se	APIC_Inte	APWS_W	APLU_Lar	APLG_list	APLA_ Ari
IELTS	1	0.043	-0.015	0.053	0.067	0.062	0.101	0.058	0.087	0.058	-0.056	-0.032	0.052	0.030	0.009	0.047	-0.001
APCL_COLL_Le	0.043	1	-0.027	0.156	0.226	0.103	0.111	0.136	0.097	0.117	0.052	0.063	0.056	0.064	0.098	0.034	0.058
Age	-0.015	-0.027	1	-0.199	-0.071	0.011	0.010	-0.008	-0.067	-0.144	-0.104	-0.087	-0.004	-0.059	-0.074	0.004	0.023
High School %	0.053	0.156	-0.199	1	0.300	0.239	0.332	0.243	0.266	0.239	0.170	0.194	0.216	0.249	0.255	0.181	0.075
Term 1_GPA	0.067	0.226	-0.071	0.300	1	0.512	0.529	0.410	0.358	0.149	0.094	0.149	0.087	0.125	0.062	0.052	0.031
Term 3_GPA	0.062	0.103	0.011	0.239	0.512	1	0.693	0.563	0.566	0.148	0.066	0.053	0.094	0.119	0.075	0.048	0.004
Earned_Hours	0.101	0.111	0.010	0.332	0.529	0.693	1	0.651	0.623	0.173	-0.031	0.036	0.049	0.082	0.081	0.068	0.054
Term 4_GPA_No	0.058	0.136	-0.008	0.243	0.410	0.563	0.651	1	0.497	0.202	0.020	0.037	0.060	0.080	0.054	0.056	0.064
Term 2_GPA	0.087	0.097	-0.067	0.266	0.358	0.566	0.623	0.497	1	0.170	0.057	0.054	0.042	0.068	0.061	0.031	0.023
APEA_ELEM_Al	0.058	0.117	-0.144	0.239	0.149	0.148	0.173	0.202	0.170	1	0.182	0.223	0.244	0.276	0.310	0.180	0.068
APRS_Reading_	-0.056	0.052	-0.104	0.170	0.094	0.066	-0.031	0.020	0.057	0.182	1	0.697	0.699	0.666	0.570	0.587	0.039
APSM_Sentence	-0.032	0.063	-0.087	0.194	0.149	0.053	0.036	0.037	0.054	0.223	0.697	1	0.676	0.810	0.590	0.593	0.177
APIC_Integ_Core	0.052	0.056	-0.004	0.216	0.087	0.094	0.049	0.060	0.042	0.244	0.699	0.676	1	0.916	0.695	0.704	0.183
APWS_Writing_W	0.030	0.064	-0.059	0.249	0.125	0.119	0.082	0.080	0.068	0.276	0.666	0.810	0.916	1	0.814	0.662	0.218
APLU_Lang_Use	0.009	0.098	-0.074	0.255	0.062	0.075	0.081	0.054	0.061	0.310	0.570	0.590	0.695	0.814	1	0.633	0.182
APLG_listening	0.047	0.034	0.004	0.181	0.052	0.048	0.068	0.056	0.031	0.180	0.587	0.593	0.704	0.662	0.633	1	0.158
APLA_ Arithmetic	-0.001	0.058	0.023	0.075	0.031	0.004	0.054	0.064	0.023	0.068	0.039	0.177	0.183	0.218	0.182	0.158	1

As seen from the above matrix there are two variables highlighted in yellow (Earned hours and Term_3 GPA) that have good correlation with the dependent variable (term_4 GPA), but it is noticed that the correlation between independent variables themselves are high for some variables in bold. So, instead of including all these variables in the model, factor analysis carried out to identify the factors and the main variables associated with each factor.

Four factors considered. For each factor variables that scored 0.7 and above KMO individual measures, variables under each factor scored 0.7 and above. For each factor the variable that has high correlation with the dependent variable considered as a representative.

Liner Regression

Running stepwise regression with 70% of the data used to train the model and 30% for testing model lead us to have only two key factors that affected the model. these are:

- Earned Hours
- Term_3 GPA.

and the equation of the model (Term 4_GPA) obtained is:

Term 4_GPA = 0.053409195203362+9.97955243795886E-03*Earned_Hours+6.12394908490219E-02*Term 3_GPA

Further details about the model are illustrated in the appendix

Regression Trees

The same process followed as previous section, and the regression trees algorithm using 'rpart' formed. 70% of the data used for training the model. The generated rules are illustrated in the appendix.

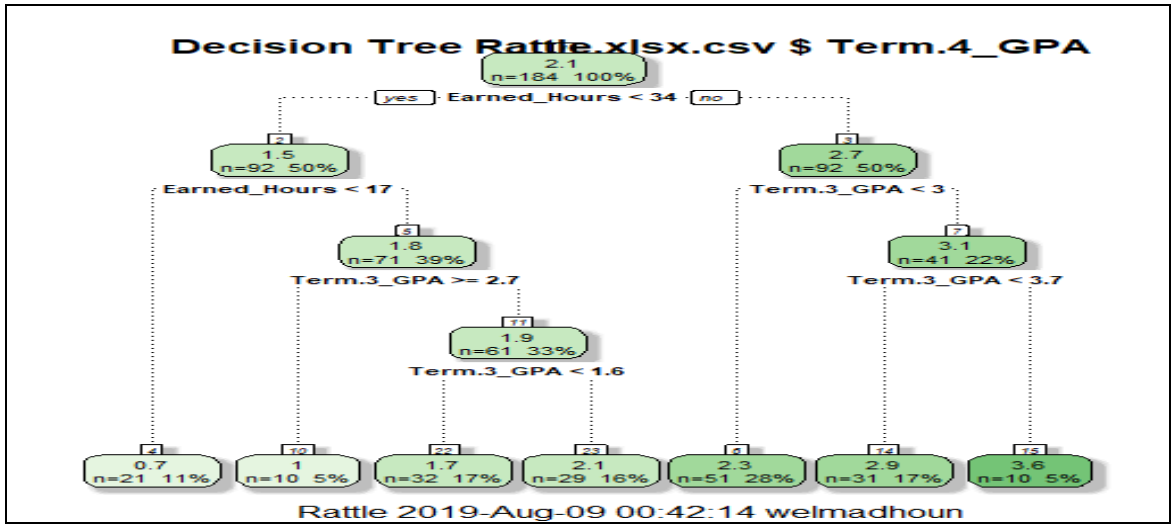


Figure 6. 17 Tree structure- students with 4_term GPA and admitted in term 2

Random Forest

Random Forest algorithm carried out as a third method to check if it will give better performance based on Pseudo R-square. The error plot created. The model started with default number of tress which is 500.

The outputs show that Mean of squared residuals is 0.770358 and the percent of variance explained is 32% and as well as the Pseudo R- Square is 0.68.

In terms of the variable importance, the result shows that Earned Hours and term_3 GPA are the most important variables. The next section will address the third group of students who have 6_term GPA and admitted in term 2.

6.4.3. Group 3: Students with 6_term GPA

The dataset of this group includes 168 students who have 6_term GPA as well as students' results in other tests and courses finished, and the number of earned hours. The historical data used to explore the possibility of predicting the GPA of term_6.

The same approach followed with testing the normality of the dependent variable (Term_6 GPA) as pre-process of the linear regression. Shapiro-Wilk test applied, and the significance level is 5%.

The null hypothesis is H0: The “Term_6 GPA” variable follows a Normal distribution while the alternative hypothesis is H1: The "Term_6 GPA" variable does not follow a Normal distribution.

Further details about the test are illustrated in the appendix.

As the computed p-value is greater than the significance level alpha = 0.05, the null hypothesis H0 can’t be rejected.

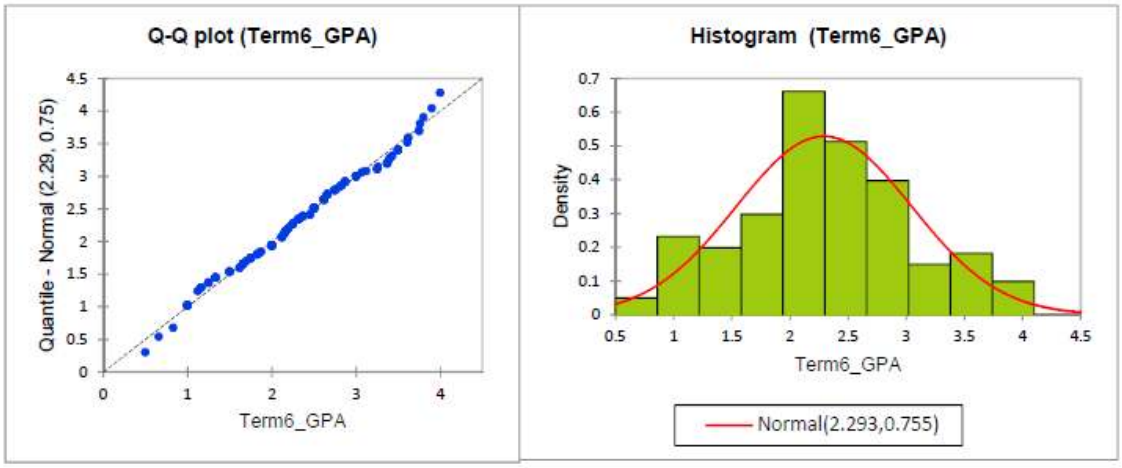


Figure 6. 18 Q-Q plot (Term6_GPA) and the second graph shows Histogram of the dependent variable

As the dependent variable is normally distributed, we will start the linear regression model and explore the possibility of predicting the GPA of term_6.

Linear Regression

Stepwise linear regression was carried out to explore the possibility of predicting term_6 GPA for students who finished six terms. The dataset of 169 students includes students results in different tests as well as their GPA in five terms. The dependent variable is term_6 GPA.

70% of the data were used to train the model, and the remaining to test the model. The process started with performing Pearson correlation to explore if there is any correlation between independent variables to identify if there is a need for factor analysis or not.

Type III Sum of Squares analysis (Term6_GPA) is illustrated below.

Table 6. 10 Type III Sum of Squares analysis (Term_6 GPA)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Age	0	0.000			
Term1_GPA	1	3.027	3.027	8.894	0.004
Term5_GPA	1	9.439	9.439	27.737	< 0.0001
Term3_GPA	1	2.628	2.628	7.722	0.007
Earned_Hours	0	0.000			
Term2_GPA	0	0.000			
Term4 GPA	0	0.000			

As seen from Type III sum of squares analysis above only three variables are contributing to the model. These are:

- Term_1 GPA
- Term_3 GPA
- Term_5 GPA

Equation of the model (Term6_GPA):

Term6_GPA =

0.852534675008481+0.156913585416397*Term1_GPA+0.343713810608238*Term5_GPA+0.18027 9075280481*Term3_GPA

Given the R-Square, 47% of the variability of the dependent variable Term6_GPA is explained by the 3 explanatory variables, and Among the explanatory variables, based on the Type III sum of squares, variable Term5_GPA is the most influential.

If we re-run the model including the above three variables to explore any improvement, the result showed that R-square remains the same as indicated below.

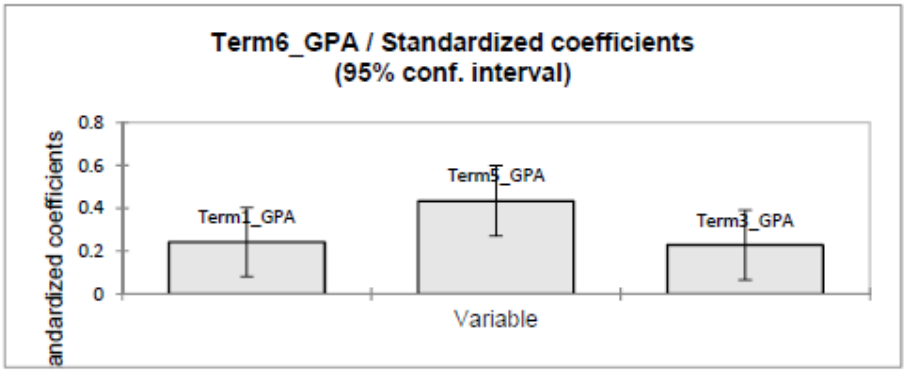


Figure 6. 19 Term_6 GPA/ Standardized coefficient including the three variables

Regression Trees

As above, the regression trees implemented. 70% of the data used to train the model.

Since from the above discussion three variables that contribute to the model identified and, the same variables used to build the tree. Since the interface of Rattle is easy to use in building up the tree, Rattle used for this process. A summary about the tree is illustrated in the appendix.

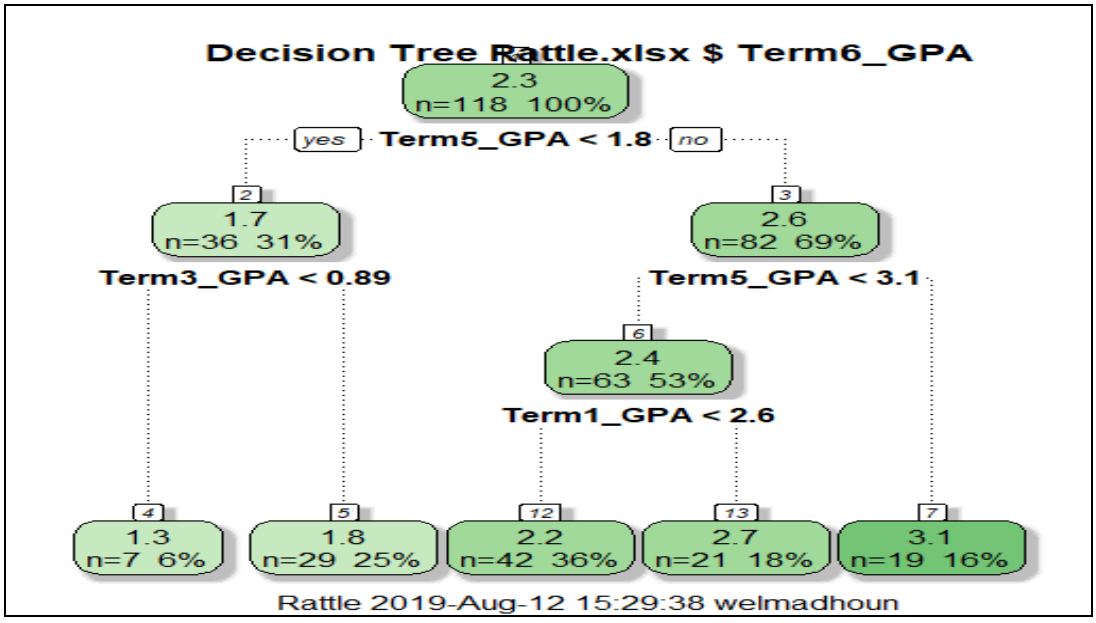


Figure 6. 20 Regression Tree- students with 6_term GPA and admitted in term 2

As we see, and as a part of evaluating the model, and based on Pseudo R-square (0.66), the independent variables explained 66% of the variability of the dependent variable.

Random Forest

Random forest was carried out in order to explore if it will give better results compared with the previous two methods. A summary of the implemented method in the appendix.

Based on the implemented method, the output showed that the most important variable is term_5 GPA as seen below.

Variable Importance

==========

	Importance
Term5_GPA	0.15445034
Term1_GPA	0.10931522
Term3_GPA	0.04717191

6.5. Evaluation of the results of the machine learning algorithms

Back to the dataset of students who joined the university in term 1 (Fall), and in particular students who have 2_term GPA, when applied the three data science algorithms, it is observed that for multiple linear regression model, Multiple R-Squared is 0.56 and adjusted R-Square is 0.56.

The analysis of variance showed that the main factors that affect students' term_2 GPA are their performance in previous term (Term_1 GPA, as well as Earned credit hours, and the performance at high school level (High school %). The regression tree showed the same conclusion as the main contributors are the number of Earned credit hours and the performance in term1 (Term_1 GPA) are the most important factors. The regression trees performed better than the linear regression models as R- Square is 0.76.

Random Forest gives Mean of Squared Error (MSE) is 0.025, and R-square is 0.75. In terms of variable important, the number of earned hours is the most important variable, and the GPA of term_1. The predictive vs. observed plots (the predicted values against the observed values with two lines) compared as well. One represents the linear fit to the actual points, and the other one for perfect fit, and for each plot, the Pseudo R-Squared that measures the square of the correlation between the predicted and observed value identified. The closer to 1, the better. This measure was compared for all applied algorithms, and when it comes to students who have 2-term_GPA, the comparison is illustrated in the table below:

Table 6. 11 Pseudo R-Squared for student with 2_term GPA

Algorithm	Pseudo R-Squared
Multiple Linear Regression	0.56
Regression Tree	0.76
Random Forest	0.75

Table 6. 12 Pseudo R-Squared for students with 4_term GPA

Algorithm	Pseudo R-Squared
Multiple Linear Regression	0.59
Regression Tree	0.64
Random Forest	0.71

As seen, Random Forests is scored higher and performed better compared with other algorithms, and the percent of variance explained 71% which is reasonable.

When it comes to the dataset of students who have 6_term GPA, the same conclusion was observed as illustrated in the comparison below.

Table 6. 13 Pseudo R-Squared for students with 6_term GPA

Algorithm	Pseudo R-Squared
Multiple Linear Regression	0.99
Regression Tree	0.94
Random Forest	0.94

From the above table, R-Square is almost the same for Random Forest and regression Trees. Since the multiple R-Square values which defined as coefficient determination is a measure of the model's performance, and how well the built model explains the dependent variable which in our case is 2_term GPA, 4_term GPA, and 6_term GPA; we can look for it as a correlation coefficient, and based on this, the closer to 1; the better the model works, and explains the variation in the dependent variable.

Given that, and based on the preceding comparison tables, we can say that for the above three algorithms, Random Forests is explained 75%, 71%, and 94% of the variation in the dependent variable which is quite good.

Analysing the outputs of the same algorithms built up for datasets related to students who joined the university in term 2 (Spring), it was concluded that the same main factors affected the performance for term 1 groups are almost the same for the three groups of students admitted in term 2. These are students' previous GPAs as well as the number of earned credit hours, without identifying any major contributions of students' performance at the Foundation Program levels. This conclusion will be addressed with further details in coming sections.

As above, and in terms of the level of Pseudo R- Square, the table below illustrates a comparison of R- Square for all algorithms as follows:

Table 6. 14 Comparison of the Pseudo R-Squared related to algorithms built for students admitted in term 1

Algorithm	Pseudo R-Squared 2 term GPA	Pseudo R-Squared 4_term GPA	Pseudo R-Squared 6_term GPA
Multiple Linear Regression	0.56	0.59	0.99
Regression Tree	0.76	0.64	0.94
Random Forest	0.75	0.71	0.94

As seen from the above table, Random forests algorithms are performing better for the first two groups compared to other algorithms.

6.6. Summary and relationship with the findings

This chapter presents the results of applying the three data science algorithms on the three groups of students who have 2_term GPA, 4_term GPA, and 6_terrm GPA, either students admitted in term 1 or term 2.

In all cases, 70-80% of the data used to train the model, and the remaining used for testing. For each algorithm, the main factors affected the student's GPA was extracted and it was noticed that in all cases, none of the Foundation Program courses add any value to the model. This was a surprise given the fact that the main objective of the Foundation Program is to prepare students for their future academic journey in the major/ college they will join.

In terms of the relationship of these findings with the preceding group (Foundation students), it is concluded that certain Mathematics and English language courses are affecting students' performance. Also, it is noted that the Regression tree performed better compare with other applied algorithms.

The situation is different when it comes to the larger group of students who declared their major and have 2_term GPA, 4_term GPA and 6_term GPA. The findings for these groups showed that only the performance in previous term as well as earned credit hours are the main contributors to the built models while none of the Foundation courses adding any value.

For the applied algorithms, the findings are almost the same as in most of the tested cases the linear regression and regression tree performed better compared with other algorithms.

As this is unexpected and important conclusion, and to better understand further details, the next chapter will address further exploration in order to identify if the Foundation Program affect the performance and students' GPA in certain college/ academic program to identify which College or academic Program specifically benefit from the Foundation Program.

Chapter 7 Testing the correlation between the student's performance in the Foundation Program per College and students' GPA(s) after declaring the major

7.1. Introduction

As mentioned in the preceding discussion, the university established the Foundation Program to bridge the gaps between high school and the university for all students admitted in Science, Health, Pharmacy, Medicine, and Engineering Colleges. Intensive Mathematics and English courses are offered over two consecutive terms to enhance students' skills in these subjects. The Foundation Program was established by the concerned university during the university's Reform Project during the period 2003- 2007. During this period, different offices and academic programs, including the Foundation Program were established. The mentioned Program has its own hiring process for academic and non- academic staff, human and financial resources, buildings, as well as annual budget dedicated to the Program annually. Students are required to pass these courses with a grade of pass/ fail. If the student fails to finish all Foundation requirements in two terms (full academic year), they can spend another two terms to meet the Program's requirements.

Historical data indicates that many students reached the maximum allowed period (Two academic years) without passing the Program's requirements. Based on this, their admission is suspended until they pass specific international Mathematics/ English tests with certain scores.

The discussion in the preceding chapter flagged an important conclusion that for students who have 2_term GPA, 4_term GPA, and 6_term GPA either who admitted in term 1 or term 2. Findings revealed that the Foundation Program doesn't add any value to their performance

during their enrolment in the Colleges. This point needs further investigation, and results might lead to major changes in the admission process for students in the program, as well as the implication of requiring students to spend two years to pass the Program's requirements. The Foundation program is a mandatory pre-entry program to all students admitted in the following colleges:

Science, Health, Engineering, Medicine, and Pharmacy

Since student must finish the Foundation program before declaring his/ her major in one of the above colleges, a Pearson correlation test was carried out to test the correlation between the student's performance in the Foundation program and the GPA after declaring the major.

The initial test was applied to all colleges based on the number of terms' GPAs students have as follows:

- Students who have 2_term_GPA,
- Students who have 4_term_GPA,
- Students who have 6-term_GPA

Results showed no correlation between the student's performance at the Foundation program and their GPA after moving to one of the above colleges.

In order to identify if any of the Colleges above benefit from the Foundation program specifically and its' results reflected on students after declaring the major, the dataset was divided per college term. GPA and correlation tests were carried out again for each college individually to test the relationship between the students' performance in the Foundation program and their GPA after declaring the major in one of these colleges. The test was carried out for those who have 2_term_GPA, 4_term_GPA, and 6_term_GPA.

This chapter addresses further investigation to study the relationship between students' performance in the Foundation Program and their GPA after they declared their major in one of the academic programs s/he joint in order to identify which students benefit from the Foundation course in their future journey in the student's chosen college.

7.2. Data Preparation and input data.

The datasets of all students affiliated with Science, Engineering, Health, Pharmacy and Medicine Colleges, including all students who passed the Foundation Program successfully, were extracted and grouped based on the College in which the student enrolled. The list of students was grouped per admission term individually despite the number of terms the student finished as the main point is to check the implication of the Foundation Program on students' GPA in general.

The datasets include either the Foundation courses results, or the international tests that the students must pass at the Foundation level. These are:

- APCL_COLL_Level_MATH
- APEA_ELEM_ALGEBRA
- APSM_Sentence_Meaning
- APRS_Reading_Skills
- APLU_Lang_Use
- APWS_Writing_Workshop
- APLG_listening
- APIC_Integ_Core
- IELTS

- ACT

7.3. Size of the data

The list of students who finished the Foundation Program was extracted, and classified based on the admission term, as well as the College in which the student enrolled. The table below illustrates the distribution for students admitted in Term 1(Fall), and passed the Foundation Program:

Table 7. 1 Number of students finished the Foundation Program per college and admitted in term 1

College students enrolled in	No. of students
Engineering	113
Health and Science	83
Pharmacy	25
Medicine	18
Total	239

The total number of students admitted in term 1 and finished the Foundation Program successfully is **239** students, for students admitted in term 2, and finished the Foundation Program, the size of the data is **609** students.

The next section will address the correlation tests that performed to identify the relationship (if any) between the Foundation courses outcomes and the student accumulative GPA in the College they join.

7.4. Correlation tests: Term 1

Pearson correlation test was performed. Each College was treated separately to examine if any specific college benefitted from the Foundation Program and has an implication on the student's

GPA. The discussion below presents the outcomes of the correlation test for each college individually.

7.4.1. Students with 2_term GPA

Pearson correlation was performed using Rattle. We used Rattle to perform the correlation test on datasets related to all colleges. The tables and charts below illustrate the outcomes of the correlation tests performed as follows:

The analysis below shows a comparison of the outcomes. For **College of Engineering**, the correlation test carried out and the results are indicated in the matrix below as follows:

Table 7. 2 Correlation Matrix (Pearson)- Engineering students admitted in term 1

Variables	APLA_ Arit	APRS_Readi	APSM_Sente	APWS_Writi	APIC_Integ	APLU_Lang_	APLG_liste	APEA_ELEM_	Student's GPA	APCL_COLL_	ACT	IELTS	SAT
APLA_ Arithmetic	1	0.192	0.201	0.191	0.191	0.166	0.130	0.159	0.167	0.016	-0.097	0.030	0.000
APRS_Reading_Skills	0.192	1	0.790	0.807	0.919	0.783	0.795	0.258	0.175	-0.009	0.113	0.015	0.000
APSM_Sentence_Meaning	0.201	0.790	1	0.981	0.947	0.897	0.801	0.221	0.191	0.046	0.108	0.039	0.000
APWS_Writing_Workshop	0.191	0.807	0.981	1	0.966	0.966	0.816	0.269	0.219	0.065	0.107	0.038	0.000
APIC_Integ_Core	0.191	0.919	0.947	0.966	1	0.935	0.898	0.264	0.199	0.033	0.127	0.034	0.000
APLU_Lang_Use	0.166	0.783	0.897	0.966	0.935	1	0.795	0.318	0.246	0.087	0.098	0.035	0.000
APLG_listening	0.130	0.795	0.801	0.816	0.898	0.795	1	0.181	0.124	-0.011	0.151	0.042	0.000
APEA_ELEM_ALGEBRA	0.159	0.258	0.221	0.269	0.264	0.318	0.181	1	0.385	0.181	-0.027	0.079	0.000
Accum_GPA	0.167	0.175	0.191	0.219	0.199	0.246	0.124	0.385	1	0.257	0.055	0.201	0.100
APCL_COLL_Level_MATH	0.016	-0.009	0.046	0.065	0.033	0.087	-0.011	0.181	0.257	1	-0.021	0.083	0.000
ACT	-0.097	0.113	0.108	0.107	0.127	0.098	0.151	-0.027	0.055	-0.021	1	0.310	0.000
IELTS	0.030	0.015	0.039	0.038	0.034	0.035	0.042	0.079	0.201	0.083	0.310	1	0.201
SAT	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.201	1

For College of **Health Sciences**, the correlation matrix is illustrated as follows:

Table 7. 3 Correlation Matrix (Pearson)- Health science students admitted in term 1ted in term 1

Variables	Term_2_GP	APCL_COLL	ACT	IELTS	Term_1_GP	APLA_ Ari	APRS_Read	APSM_Sent	APWS_Writ	APIC_Inte	APLU_Lang	APLG_list	APEA_ELEM	Student's GPA	SAT
Term_2_GPA	1	0.078	-0.002	-0.026	0.100	0.062	0.006	0.017	0.045	0.041	0.058	0.077	0.007	0.016	0.000
APCL_COLL_Leve	0.078	1	0.209	0.092	0.064	-0.017	-0.023	-0.018	-0.043	-0.043	-0.048	-0.066	0.086	0.207	0.000
ACT	-0.002	0.209	1	0.401	0.042	-0.001	0.029	0.008	0.023	0.012	0.035	-0.060	0.023	0.086	0.000
IELTS	-0.026	0.092	0.401	1	-0.144	-0.007	-0.092	-0.058	-0.085	-0.090	-0.101	-0.064	0.132	0.209	-0.047
Term_1_GPA	0.100	0.064	0.042	-0.144	1	0.082	0.170	0.247	0.307	0.251	0.331	0.138	0.088	0.181	0.000
APLA_ Arithmet	0.062	-0.017	-0.001	-0.007	0.082	1	0.314	0.259	0.290	0.300	0.286	0.188	0.293	0.220	0.000
APRS_Reading_S	0.006	-0.023	0.029	-0.092	0.170	0.314	1	0.779	0.820	0.915	0.754	0.707	0.396	0.330	0.000
APSM_Sentence_	0.017	-0.018	0.008	-0.058	0.247	0.259	0.779	1	0.937	0.912	0.755	0.747	0.330	0.169	0.000
APWS_Writing_W	0.045	-0.043	0.023	-0.085	0.307	0.290	0.820	0.937	1	0.975	0.935	0.829	0.381	0.201	0.000
APIC_Integ_Cor	0.041	-0.043	0.012	-0.090	0.251	0.300	0.915	0.912	0.975	1	0.911	0.870	0.389	0.240	0.000
APLU_Lang_Use	0.058	-0.048	0.035	-0.101	0.331	0.286	0.754	0.755	0.935	0.911	1	0.797	0.401	0.207	0.000
APLG_listening	0.077	-0.066	-0.060	-0.064	0.138	0.188	0.707	0.747	0.829	0.870	0.797	1	0.258	0.095	0.000
APEA_ELEM_ALGE	0.007	0.086	0.023	0.132	0.088	0.293	0.396	0.330	0.381	0.389	0.401	0.258	1	0.543	0.000
Accum_GPA	0.016	0.207	0.086	0.209	0.181	0.220	0.330	0.169	0.201	0.240	0.207	0.095	0.543	1	-0.069
SAT	0.000	0.000	0.000	-0.047	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.069	1

For the College of **Pharmacy**, the outputs are indicated in the table below as seen.

Table 7. 4 Correlation Matrix (Pearson)- Pharmacy students with 2_term GPA and admitted in term 1

Variables	SAT	ACT	IELTS	Student's GPA	COLL_Level	ELEM_ALG	Reading_S	entence_M	riting_Wo	C_Integ_C	LG_listeni	LU_Lang_U
SAT	1	0.000	0.130	-0.291	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ACT	0.000	1	0.162	0.087	0.028	0.004	0.000	0.000	0.000	0.000	0.000	0.000
IELTS	0.130	0.162	1	0.285	0.152	-0.140	-0.088	-0.024	0.010	-0.030	0.000	0.037
Accum_GPA	-0.291	0.087	0.285	1	0.706	0.626	0.328	0.405	0.452	0.411	0.417	0.472
APCL_COLL_Level_MA	0.000	0.028	0.152	0.706	1	0.791	0.550	0.605	0.593	0.580	0.516	0.559
APEA_ELEM_ALGEBRA	0.000	0.004	-0.140	0.626	0.791	1	0.763	0.728	0.694	0.745	0.711	0.638
APRS_Reading_Skill	0.000	0.000	-0.088	0.328	0.550	0.763	1	0.955	0.898	0.965	0.897	0.817
APSM_Sentence_Mear	0.000	0.000	-0.024	0.405	0.605	0.728	0.955	1	0.972	0.983	0.909	0.909
APWS_Writing_Works	0.000	0.000	0.010	0.452	0.593	0.694	0.898	0.972	1	0.979	0.938	0.982
APIC_Integ_Core	0.000	0.000	-0.030	0.411	0.580	0.745	0.965	0.983	0.979	1	0.962	0.936
APLG_listening	0.000	0.000	0.000	0.417	0.516	0.711	0.897	0.909	0.938	0.962	1	0.922
APLU_Lang_Use	0.000	0.000	0.037	0.472	0.559	0.638	0.817	0.909	0.982	0.936	0.922	1

For **College of Medicine**, we performed the test as well and the output matrix is illustrated below as follows:

Table 7. 5 Correlation Matrix (Pearson)- College of Medicine and admitted in term 1

Variables	SAT	Student's GPA	APCL_COLL	IELTS	ACT	APLA_ Ari	APRS_Read	APLG_list	APIC_Inte	APSM_Sent	APWS_Writ	APLU_Lang	APEA_ELEM_ALGEBRA
SAT	1	-0.548	0.000	0.133	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Accum_GPA	-0.548	1	0.346	-0.204	0.193	-0.009	0.058	0.125	0.137	0.073	0.155	0.188	0.022
APCL_COLL_Level_MA	0.000	0.346	1	0.285	0.000	0.000	0.070	0.010	0.000	-0.063	-0.047	-0.030	0.135
IELTS	0.133	-0.204	0.285	1	0.594	0.178	0.130	0.126	0.084	0.071	-0.010	-0.060	0.158
ACT	0.000	0.193	0.000	0.594	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
APLA_ Arithmetic	0.000	-0.009	0.000	0.178	0.000	1	0.517	0.593	0.541	0.288	0.349	0.346	0.069
APRS_Reading_Skill	0.000	0.058	0.070	0.130	0.000	0.517	1	0.869	0.840	0.521	0.501	0.429	-0.169
APLG_listening	0.000	0.125	0.010	0.126	0.000	0.593	0.869	1	0.958	0.674	0.696	0.627	0.083
APIC_Integ_Core	0.000	0.137	0.000	0.084	0.000	0.541	0.840	0.958	1	0.806	0.861	0.793	0.089
APSM_Sentence_Mear	0.000	0.073	-0.063	0.071	0.000	0.288	0.521	0.674	0.806	1	0.884	0.705	0.135
APWS_Writing_Works	0.000	0.155	-0.047	-0.010	0.000	0.349	0.501	0.696	0.861	0.884	1	0.954	0.199
APLU_Lang_Use	0.000	0.188	-0.030	-0.060	0.000	0.346	0.429	0.627	0.793	0.705	0.954	1	0.217
APEA_ELEM_ALGEBRA	0.000	0.022	0.135	0.158	0.000	0.069	-0.169	0.083	0.089	0.135	0.199	0.217	1

Analysing the correlation matrix for all students who have 2_term GPA, it has been noted that for all Colleges except College of **Pharmacy**, there is no significant correlation between the students' GPA and the courses they finished at the Foundation level.

For students who enrol in the College of **Pharmacy,** and have 2_term GPA, it was noticed that there is a correlation (0.62) between students' GPA, and the Mathematics course (APEA_ Elementary Algebra) as well as the course “APCL_Coll_ Level Math.”. Also, there is a correlation (0.70) between the GPA and the student's performance in the English Foundation courses “APSM_ Sentence _Meaning," and “APRS_Reading skills” (0.71). For all colleges as mentioned above, the same test was carried out for students who have a 2_term GPA and declared their majors after they passed the Foundation Program successfully.

7.4.2 Students with 4_term GPA

The same correlation test was carried out for all students enrolled in each College, and as above, it was noticed that there is no significant correlation between the students' GPA and his/ her performance at the Foundation level despite the College the student enrols in.

- **Students with 6_term GPA**

The same test was carried out for this group as well, and as above, and results showed no significant correlation between students' GPA and their performance at the Foundation level. Although, for students enrolled in College of Pharmacy as the test showed that there is a high correlation (0.90) between the student's GPA and the performance in the English courses that were offered at the Foundation Program. A good example of this result is the course “APIC_ Integ_Core," and for the course "APLG_ Listening ACT," the correlation is (0.90). The same thing for the course “APLU_ Lang.Use”, the correlation is 0.70.

To confirm the above conclusion, and data revealed that for all students, (2_term GPA, 4_term GPA or 6_term GPA), there is no relationship between Foundation courses and the Undergraduate GPA they earned after declaring the major except the College of Pharmacy who have 2_ term GPA and 6_term GPA. The table below shows the correlation scores for the College of Pharmacy as follows:

Table 7. 6 Correlation Matrix (Pearson)- Pharmacy students with 6_term GPA and admitted in term 1

Variables	ELEM_ALG	ACT	Reading_S	Student's GPA	LG_listeni	riting_Wo	C_Integ_C	LU_Lang_U	entence_M	COLL_Level	IELTS
APEA_ELEM_ALGEBRA	1	−0.234	−0.301	−0.201	0.148	0.481	0.229	0.313	0.523	−0.106	−0.405
ACT	−0.234	1	0.866	0.901	0.720	0.674	0.771	0.475	0.706	−0.500	0.577
APRS_Reading_Skills	−0.301	0.866	1	0.991	0.891	0.682	0.859	0.742	0.531	0.000	0.151
Accum_GPA	−0.201	0.901	0.991	1	0.918	0.761	0.906	0.764	0.637	−0.092	0.185
APLG_listening	0.148	0.720	0.891	0.918	1	0.906	0.985	0.953	0.728	0.078	−0.146
APWS_Writing_Worksho	0.481	0.674	0.682	0.761	0.906	1	0.958	0.865	0.935	−0.206	−0.053
APIC_Integ_Core	0.229	0.771	0.859	0.906	0.985	0.958	1	0.915	0.835	−0.084	−0.038
APLU_Lang_Use	0.313	0.475	0.742	0.764	0.953	0.865	0.915	1	0.631	0.300	−0.433
APSM_Sentence_Meanin	0.523	0.706	0.531	0.637	0.728	0.935	0.835	0.631	1	−0.530	0.224
APCL_COLL_Level_MATH	−0.106	−0.500	0.000	−0.092	0.078	−0.206	−0.084	0.300	−0.530	1	−0.866
IELTS	−0.405	0.577	0.151	0.185	−0.146	−0.053	−0.038	−0.433	0.224	−0.866	1

Correlation tests: Term 2

The same analysis was performed on the datasets of students admitted in term 2 (spring) and passed the Foundation Program to explore the implication of the Foundation Program on their performance after declaring their majors. As we see from the above matrix, the highest correlation is between the student's GPA and the mathematics course named "APCL_Coll_Level_MATH."

Table 7. 7 Correlation Matrix- Pharmacy students admitted in term 2

Variables	APIC_Inte	APWS_Writ	APLU_Lang	APSM_Sent	APLG_list	APRS_Read	High_Scho	APL_Accup	APEA_ELEM	Student's GPA	APLA_ Ari	Earned_Ho	IELTS	APCL_COLL_Level_MATH
APIC_Integ_Core	1	0.942	0.923	0.886	0.917	0.917	-0.347	0.143	0.186	0.215	-0.190	-0.191	0.163	0.094
APWS_Writing_Works	0.942	1	0.950	0.947	0.789	0.744	-0.327	-0.119	-0.002	0.205	-0.231	-0.135	0.095	0.085
APLU_Lang_Use	0.923	0.950	1	0.839	0.794	0.762	-0.309	0.112	0.145	0.154	-0.095	-0.059	0.103	0.139
APSM_Sentence_Mean	0.886	0.947	0.839	1	0.728	0.699	-0.336	-0.167	-0.137	0.128	-0.238	-0.035	0.091	0.114
APLG_listening	0.917	0.789	0.794	0.728	1	0.862	-0.417	0.267	0.261	0.212	-0.156	-0.196	0.195	0.005
APRS_Reading_Skill	0.917	0.744	0.762	0.699	0.862	1	-0.263	0.417	0.371	0.187	-0.111	-0.227	0.208	0.136
High_School_%	-0.347	-0.327	-0.309	-0.336	-0.417	-0.263	1	0.013	0.183	0.456	0.222	-0.103	0.252	0.506
APL_Accuplacer	0.143	-0.119	0.112	-0.167	0.267	0.417	0.013	1	0.595	-0.142	0.334	0.098	0.095	0.197
APEA_ELEM_ALGEBRA	0.186	-0.002	0.145	-0.137	0.261	0.371	0.183	0.595	1	0.289	0.508	-0.028	0.318	0.215
Accum_GPA	0.215	0.205	0.154	0.128	0.212	0.187	0.456	-0.142	0.289	1	0.040	-0.236	0.537	0.329
APLA_ Arithmetic	-0.190	-0.231	-0.095	-0.238	-0.156	-0.111	0.222	0.334	0.508	0.040	1	0.609	0.059	0.252
Earned_Hours	-0.191	-0.135	-0.059	-0.035	-0.196	-0.227	-0.103	0.098	-0.028	-0.236	0.609	1	-0.098	0.395
IELTS	0.163	0.095	0.103	0.091	0.195	0.208	0.252	0.095	0.318	0.537	0.059	-0.098	1	0.437
APCL_COLL_Level_MA	0.094	0.085	0.139	0.114	0.005	0.136	0.506	0.197	0.215	0.329	0.252	0.395	0.437	1

7.5 Summary

This chapter discussed the correlation between students' performance enrolled in each college and their performance in the Foundation mathematics and English courses they finished at the Foundation level. The result showed that there is no significant correlation between students' GPA and the performance at the Foundation level except for students enrol in the College of Pharmacy. This conclusion was clear for students admitted either in term 1 or term 2.

Chapter 8 Discussion and Conclusions

8.1 Introduction

This chapter summarizes the results and their relationship to the research questions, the need of the study as identified in the introduction, the significance of the results and factors that affect students' academic performance at different levels are discussed. Finally, an elaboration of the limitations and areas for future research work are presented.

8.2. Achieved Objectives

The aim of this research was to explore the possibility of building predictive models that help academic institutes predict students' performance and the student's GPA. This will help in identifying at-risk students who have a low GPA (below 2.0 out of 4.0) and were placed on academic probation during their educational journey.

To achieve this goal, students were grouped according to their status, and the academic requirements that affect their performance. Several models were built to classify and predict the GPA for each category. The built models helped in identifying the main factors that could affect the student's GPA.

8.3 Discussion and Findings

The main objectives for this work are to explore the possibility of building predictive models to predict students' performance at an early stage, prevent students from being dismissed from the university and identifying the factor(s) that play significant role in the student's GPA. The above work focused on building two models for students enrolled in the Foundation Program.

Variables included are the admission term, undergraduate program, passing grade for the Foundation Program or those who enrolled in the Colleges that don't require the Foundation Program. For each category, either Foundation level or Undergraduate level, the discussion is split into two subsections. These are:

- Findings related to data science algorithms applied, and
- Findings related to the factors that affect students' performance.

8.3.1. Foundation Level

For the Foundation Program, the size of the data is 769 students in total.

In terms of the data science algorithms: The Logistic Regression models were carried out on the list of students who are not exempted from any course in the Foundation Program, and their academic standing by the end of the program either pass or fail.

For students enrolling in term 1, 63% of them were female and the remaining 32% were male. The built model for this category managed to classify 88.4% correctly.

In terms of the factors that affect the academic status of this group, there are two variables. These are:

- School origin
- APLU_Language Use

School Origin refers to the high school from which the student came, private, International or local school. APLU-Language Use is an English course required at the Foundation level.

English Language level is the main problem for students who are not exempted from the Foundation Program. Students who are proficient in the English Language, are most likely to be exempted from taking Foundation English courses. If we compared this conclusion with the

variables that affected the performance for the same category, those who were admitted in **term 2**, we would notice that the variables that influenced the performance are:

- Student's age, and
- Students' performance in the Mathematics course offered by the Foundation Program.

Age is an expected factor for students admitted in term 2, as some students, especially male students, delay their admission, and give priority to work as a full-time employee in the government's ministries and companies. Usually, their admission is postponed, and it takes place in term 2 (spring).

Further investigation has been done for students partially exempted from the Foundation Program, either those who enrol in the university in term 1or term 2. As mentioned before, partially exempted means that the student was able to pass certain international Math/ English tests with specific accepted scores that allows him/ her to waive the Foundation course(s) and had the chance to register in undergraduate classes with a letter grade (A, B+, B…etc). Since the number of students in this category is small, they were excluded, and the research focused on the dataset of students enrolled in the Foundation without exemption.

For this group, students admitted in term 2 and not exempted from the Foundation Program, the built model explained only 73.3% of the variability in the dependent variable, and after trying other data science algorithms to get the results improved, it was concluded that Regression Trees were more useful. The most importance variable was the students' performance in the Foundation Math. Course Elementary Algebra named “APEA-ELEM-ALGEBRA". This result is consistent with the results from the Multiple Linear regression model.

8.3.2. Undergraduate Level

The second part of this project was focused on the other category of the dataset that related to students enrolled in the undergraduate academic programs having either already passed the Foundation program and declared their major in one of the colleges' programs, or enrolled in colleges that don't require the Foundation program.

The GPA is counted by the end of each term as these students are taking undergraduate courses with certain credit hours and get a letter grade for each session. The author treated those who admitted in term 1 and term 2 separately, and for each term, we grouped students in three groups according to the number of terms they finished. These groups are students who have 2_term GPA, 4_term GPA, and 6_term GPA.

- In terms of the data science algorithms: for the dataset of students with 2_term GPA, and among the three data science algorithms; Regression Tree and Random Forests performed better compared with the linear regression method, and For students with 4_term GPA, the three algorithms, Random Forests performed better (71%) compared with others. While when we applied the three algorithms on the dataset of students with 6_term GPA, the three algorithms gave high result with R-square (94%) and slightly higher for the regression model.
- In terms of the factors that affect the academic status, for students with 2_term GPA, the number of earned hours, Term_1 GPA and high school result are the most contributors. The same conclusion reached for students admitted in term 2. The results showed that students' GPA is not correlated or affected either positively or negatively with students' performance at the Foundation courses. This conclusion is a bit surprised as the main

purpose of initiating the Foundation Program was to bridge the gap between schools and the university and prepare students to continue their college study smoothly. When we applied the multiple linear regression on students admitted in term 1 and have 2_term GPA the previous GPA (term_1 GPA) based on students' historical record, it was noticed that term_2 GPA is mainly correlated with term_1 GPA, earned hours, and high school result. It never depends on the performance at the Foundation level. From the other hand, for those who admitted in term 2, the student's performance at high school which is used as criteria for the admission is the main factor that affected term_2 GPA. This conclusion is consistent with the fact that usually the admission is competitive and depends on students' high school result. For those who didn't get a chance to join the university in term 1, they apply again to join in term 2. So, it is normal that the GPA for this group is relying on high school result.

The same process was applied for students who have a 4_term GPA to predict term _4 GPA based on the GPA of the previous three terms. This group of students is mainly female (87%), and the remaining 13% are male. 95% of them came from local public schools; the other 5% were in an International school. The main conclusion we reached that the previous GPAs and the number of earned hours are the primary influence and contribute to the model.

It is noticed that the main factors that have an impact on students' term _ 4 GPA are:

- The number of earned hours
- Students' performance at High school (High school %), and
- The previous term GPA (term_3 GPA)

For students who have 6_term GPA and to build up the models, we used the previous five GPAs to predict the GPA of term_6. The conclusion we reached is the same as other models

as term_6 GPA is mainly affected by the earlier GPAs, as well as the number of earned hours. The same conclusion we reached for students admitted in term 2.

8.3.3. Foundation Program and its impact on students' GPA

As the main findings we reached is that no significant impact of the Foundation Program on students' GPA, and to get better insight this fact, the last part of this work focused on further exploration to identify if the Foundation Program add any value to certain college's students.

Based on this, all students finished the Foundation Program were grouped according to the College that the student affiliated to. These colleges are: Health, Science, Pharmacy, Medicine, and Engineering. For each college, Pearson correlation test was performed to test the correlation between the GPA and all Foundation courses either Mathematics or English with all levels.

The test carried out on all students passed the Foundation Program either they admitted in term 1 or term 2. For students who joined the university in term 1 and have 2_term GPA, the correlation test showed that:

- For students enrolled in the Colleges of Engineering and Health Sciences, data revealed the highest correlation between the student's GPA and Math. Foundation course named "APEA_ELEM_ALGEBRA" and it is scored 0.37 for Engineering students and 0.53 for Health Science students. Although highest correlation is a little bit low there is no other significant correlation with other courses.
- For the College of Pharmacy, the situation is different as there is a significant correlation between the GPA and below Foundation courses. The table below illustrates the correlation scores as follows:

Table 8. 1 Correlation test – students GPA and their Performance in Foundation for Pharmacy students who have 2_term GPA

Variables	**Pharmacy- GPA Correlation score by using Rattle**
APEA_ELEM_ALGEBRA	0.7534386
APSM_Sentence_Meaning	0.7389514
APRS_Reading_Skills	0.7107684
APWS_Writing_Workshop	0.6592687
APCL_COLL_Level_MATH	0.7603584
APIC_Integ_Core	0.6309739
APLU_Lang_Use	0.6106958

As seen from the above table, there is a good correlation between the GPA and most of the English/ Mathematics courses.

For the same college, and students with 6_term GPA, the correlation was much clearer. The correlation test was performed, and it is scored (0.90) between students' GPA and the Foundation English courses such as APIC Integ- Core, and APLG- Listening, and scored (0.99) with PRS_Reading_Skills.

Table 8. 2 Correlation test – students GPA and their performance in Foundation for Pharmacy students who have 6_term GPA

Variables	Student_GPA
APCL_COLL_Level_MATH	-0.092
APSM_Sentence_Meaning	0.637
APWS_Writing_Workshop	0.761
APIC_Integ_Core	0.906
APLU_Lang_Use	0.764
APLG_listening	0.918
APRS_Reading_Skills	0.991
ACT	0.901

As seen from the above table, for students enrolled in the College of Pharmacy and a 2_term GPA, their performance is affected by both Mathematics and English courses in the Foundation Program. The correlation for both classes with all levels is between 0.61- 0.76, but if the student within the same college reached year 4 (in another word, finished 6 terms, it looks like the GPA is mainly affected by the Foundation English courses only. A possible explanation is that for senior students in year 4 there is a need for English professional level as students need to read/ write in English much more than in year 1. They have to enrol in internship courses and practical training in the hospital, as well as the level of the senior courses they take in year 4. For the impact of the Foundation Program on Pharmacy students, the data seems to indicate that the impact is from the Foundation courses, but in reality within the College of Pharmacy, there is an intensive follow up, mentoring, as well as an academic advising from day 1 that is reflected in student performance. It is not only the Foundation that affects student performance in the College of Pharmacy. This is an area for further investigation.

In terms of students admitted in term 2 (spring), and after testing the correlation between the GPA and all Foundation courses, data revealed no major impact of the Foundation on the student's GPA except for students who have 2_term GPA. There is a fair correlation (0.51) between the GPA and mathematics course offered by the Foundation Program named" APCL_CoLL_Level_MATG".

8.4 Contribution

8.4.1. Question 1: Which data mining algorithm(s) is the most appropriate and effective in developing predictive model to predict students' GPA?

In general, comparing the three methods used, Random Forests performed better when compared with all other methods. Multiple Linear regression gave good results for some of the

built models. Although two different sources were used to build the models R, Rattle (R Analytical Tool To Learn Easily) and XLSTAT, the results are almost the same in most of the cases. In terms of the capabilities of these two systems in building Machine Learning algorithms and displaying the results, both are able to do the job. One of the main differences is that R and Rattle are free sources while XLSTAT is not. Although XLSTAT is not free, it is inexpensive, and is capable of handling machine learning techniques as well as evaluating the built models. One new and important feature that added to XLSTAT is its integration with R and based on this many R packages are included in XLSTAT new version.

8.4.2. Question 2: What are the main factors that affect students' GPA in each academic year?

Students enrolled in the Foundation Program

For students admitted in term 1: If they are not exempted from the Foundation Program, School origin and the student's performance in the course APLU_Language Use are the most influential factors affecting student performance. For students admitted in term 2 and not exempted from the Foundation Program, student's age and student's performance in the Foundation course " Elementary Algebra" are the most influential.

Undergraduate Level

In general, for both admission terms, and for all colleges, the main influential factors are the previous GPA and the number of earned hours despite the number of terms that the student finished except the College of Pharmacy, it was found that the student's GPA at the College of Pharmacy is affected by the students' performance in mathematics and English courses at the Foundation level, and this is the only college that affected by the Foundation level although it is a mandatory for all Science, Health, Engineering, Pharmacy and Medicine colleges.

Finally, the above built models including all stages, levels of students, and all considered groups to explore the possibility of building models to predict students' performance at early stage and before they get at risk.

8.5. Generalization of findings vs gaps and the significance of the study

This research addressed the performance of students over each academic year instead of the end of first year and graduation GPA only. Previous work addressed the student's performance in certain course(s). Although the previous work (Jiawei & Micheline, 2006, Veeramuthu *et al.* (2014) ,Tair and El Halees (2012), Vandamme *et al.* (2007), and Crooks (2009)) addressed the same topic, but most of the work applied the clustering techniques to cluster students according to test scores/ performance in year 1. The present work addressed the academic and some of non-academic variables including students' age, and gender.

Back to the literature review, it was concluded that there is a need for future research work to build models by including more variables (academic/ non-academic) to make better prediction. As an example, Asif, Merceron, and Pathan (2014) made an analysis of the performance data that was provided. This was done with the help of classification logarithm that was named as ID3, it used to identify the overall marks of the student at the end of every semester. It was fully aimed at helping both lectures and students improve on the general performance in their respective responsibilities. With clear rules set for the analysis of the data but only 50 students were approached, and the required data was gathered from them.

Data mining techniques were studied to establish their influence towards the students' performance. 151 instances of data were used from a database management system course that was held at the same concerned university. The data used was from both the academic records and from the personal records of the students. The techniques that were employed by the author include classification, outlier detection, clustering and association rules. The results that produced showed important information from both the classification models and association rules. More so, clustered students' data was also used to detect the outliers and identify their

characteristics. The overall knowledge that gathered from the study helped to improve on the students' performance in the overall database course.

In addition, Hamoud, Hashim and Awadh (2018) used data mining techniques on the educational data sets to compare their performance and later identify the best technique that can be integrated in the electronic learning web miner tool. Data were gathered from a course entitled "introduction to multimedia methods" that is offered in the 3 years that the study was performed at Qatar university. The study revealed that the accuracy and the performance of the techniques used depend on the size and the type of attributes of the dataset. The other comparison for these data logarithms is found in Romero et al (2008). In this case, the study aimed at classifying students with similar marks into various groups depending on the kind of activities that are performed in the web-based course.

The activities include, amount of the quizzes taken, assignments done, time used to attempt the quiz and the assignments among many other activities. The data set consists of 438 Qatar university students in seven different courses. The performance of the logarithm was based on the numerical rebalanced data, categorical, and numerical attributes. CART and C4.5 were found to be the best in regard to categorical data.

In Hamoud, Hashim, and Awadh (2018), classification was used to predict the success of the student and this was based on the socio-demographic variables like (ethnicity, work status, disability, education, and age) and the study environment (course block and the study program). The dataset is made up of students' data for information systems from open polytechnic in New Zealand. The classification trees include QUEST, CART, CHAID and the exhaustive CHAID. CART was later discovered to be the most effective in the study at a classification percentage

of 60.5%. Just like most of the studies performed by different scholars, the study also aims at using the tree classification method in the determination of the students' final grade. To perform the experiment, it is important to fully understand the domain where the data is to be collected; the study plan and the courses that are lectured in the department of information and technology for the female students at the college of Engineering in Qatar University. The course runs for four years with each year having two semesters of study. As it is in many programs, the first year is for introduction in both social aspects and academic aspects to be specific, courses like English, communication, religion and mathematics are introductory for this program. By the first semester of the second year, students are now expected to have started taking specialized courses coupled with a few general courses. It is mandatory for students to study 16 specialized courses and then choose 7 electives from the choices they are given; these requirements are pre-requisites to graduation. Due to the domination of the mandatory courses in the study plan, major focus will be put on these 16 specialized courses. Consequently, these mandatory courses have a big impact on the graduation grade of the student.

As seen from the above discussion, most of the previous work focused on deploying data mining techniques to predict students' performance in either in certain courses or certain group of students whereas the present work considered a generalized models to a wider population and on an institutional level, helping the higher educational institutions improve students' behaviours and skills, and thereby increase student retention.

Moreover, the previous work showed a focus on certain regression methods but with more focus on clustering techniques while this work focused on deploying three data mining algorithms to predict students' performance not just to cluster them. In sum, this work took into

consideration the institutional wise and a generalized model could be deployed on other institutions that have the same educational system.

Finally, with a combination of the algorithms, and models built at each stage and level for students who have two term GPA, four term GPA and six term GPA this research work has been able to fill in the gap mentioned in section 2.8 in building generalized model that could be used on institutional wise not just for certain group of students. By deploying these models policymakers can also use the data mining tool to identify institutions that face a high risk of student drop out and formulate the policy to assist the institutions in improving the students' retention rates.

8. 6. Limitation and Future Research Work

Some limitations of this research and areas for further research are as follows:

The datasets used in building the predictive models are mainly related to students' academic records including their GPAs per term, international test scores, as well as their results in all Foundation Program courses. Although some non-academic variables were tested, such as age and gender there are few details about students' non- academic status including students' employment. As time at work may take away time from studies, there is a need for further exploration to identify if there is any relationship between students' employment status and their academic performance and their GPAs.

The concerned university enrols significantly more females than males. That fact itself is worthy of further investigation. This study did not discuss any disparities in student

performance based on gender. This would be of interest, given the traditional role of women versus men and potential differences in pre-university preparations. It would also be interesting to examine success in the Foundation by gender. Cultural expectations of gender behaviour may result in either gender not feeling comfortable asking for academic help when necessary.

This investigation was limited to one university. An expansion into other Middle Eastern institutes may be warranted to assess similarities and differences in outcomes. These similarities or differences may be related to a variety of demographic and academic factors. Another approach in identifying the factors that affect students' academic performance is studying the impact of online management systems that help students access different learning resources at any time without the restrictions of a traditional classroom.

Furthermore, other variables could be tested to identify if there is any relationship with the academic performance and other variables such as:

- Is the institution/ major being the first choice to the student?
- Students' satisfaction
- Tuition and other costs associated with study

Characteristics of the institution were not variables in this study. Administrative structure, flexibility, adaptiveness are all areas for future study. Resources and availability, especially for working students may also be of interest. An extension of this work may be to what extent the Higher Education Administrators and management team accept the decision made by the data science algorithms in adopting changes in the educational system either on schools or Higher Education levels.

Faculty characteristics may also be of interest to future researchers. Factors such as gender, comfortableness with online Learning Management Systems, availability (online or in-person office hours).

Finally, as this research showed very little benefit from the Foundation Program in terms of student performance, it would be of interest to examine any institutional evaluations of the program. This may include outcome data on pass/fail grades, student course evaluations, student faculty evaluations, utilization of program resources by faculty/by students. If indeed the Foundation Program contributes little to the ultimate academic success of the student, it is reasonable to question the future of the program. Is the expenditure of resources and time for both the university and the student worth the result? This research provides a basis for the further examination of that question.

References

Abu-Oda, G.S. And El-Halees, A.M., 2015. Data Mining In Higher Education: University Student Dropout Case Study. *International Journal Of Data Mining & Knowledge Management Process*, *5*(1), P. 15.

Aljahani, O. 2016. A review of the contemporary international literature on student retention in higher education. International Journal of Education and Literacy Studies 4 (1): 40–52

Asif, R., Merceron, A. And Pathan, M.K., 2014. Predicting Student Academic Performance At Degree Level: A Case Study. *International Journal Of Intelligent Systems And Applications*, *7*(1), P. 49.

Delavari, N., Shirazi, M.R.A. And Beikzadeh, M.R., 2004. A New Model for Using Data Mining Technology In Higher Educational Systems. In *Information Technology Based Higher Education and Training, 2004. ITHET 2004. Proceedings of The Fifth International Conference On* (Pp. 319-324). IEEE.

Dogan, B. And Camurcu, A.Y., 2008. Association Rule Mining from An Intelligent Tutor. *Journal Of Educational Technology Systems*, *36*(4), Pp. 433-447.

Edwards, G. ,2018. Machine Learning: An introduction. https://towardsdatascience.com/machine-learning-an-introduction23b84d51e6d0

Guarín, C.E.L., Guzmán, E.L. And González, F.A., 2015. A Model To Predict Low Academic Performance At A Specific Enrolment Using Data Mining. *IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje*, *10*(3), Pp. 119-125.

Hagahmed, S. "The Impact of Academic Advising on the Retention of First-year Students in a Gulf-Arab University" (2014). Public Access Theses and Dissertations from the College of Education and Human Sciences. 219. http://digitalcommons.unl.edu/cehsdiss/219

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. 2018. Predicting Student Performance In Higher Education Institutions Using Decision Tree Analysis. *International Journal Of Interactive Multimedia And Artificial Intelligence,* 5(2), 26. Doi:10.9781/Ijimai.2018.02.004

Han, J., Pei, J. And Kamber, M., 2011. Data Mining: Concepts And Techniques. Elsevier.

Huebner, R.A., 2013. A Survey Of Educational Data-Mining Research. *Research In Higher Education Journal*, *19*

Hung, J.L. And Crooks, S.M., 2009. Examining Online Learning Patterns With Data Mining Techniques In Peer-Moderated And Teacher-Moderated Courses. *Journal Of Educational Computing Research*, *40*(2), Pp. 183-210.

Hussain, S. And Hazarika, G.C., 2014. Educational Data Mining Model Using Rattle. *Editorial Preface*, *5*(6).

Ibrahim, Zaidah & Rusli, Daliela. 2007. Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree And Linear Regression. 21st Annual SAS Malaysia Forum.

Ismail, S., 2015. Design And Implementation Of An Intelligent System To Predict The Student Graduation AGPA. *Australian Educational Computing*, *30*(2).

Jiawei H. and Micheline, K. 2006, Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

Kamath, R.S. And Kamat, R.K., 2016. *Educational Data Mining With R And Rattle*. River Publishers.

Khan, I.A. And Choi, J.T., 2014. An Application Of Educational Data Mining (EDM) Technique For Scholarship Prediction. *International Journal Of Software Engineering And Its Applications*, *8*(12), Pp. 31-42.

Kotsiantis, S. B. 2012. Use Of Machine Learning Techniques For Educational Proposes: A Decision Support System For Forecasting Students' Grades. *The Artificial Intelligence Review,* 37(4), 331-344.

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. 2004. Predicting Students' Performance In Distance Learning Using Machine Learning Techniques. *Applied Artificial Intelligence*, 18(5), 411–426.

Kovacic, Z., 2010. Early Prediction Of Student Success: Mining Students' Enrolment Data.

Kumar, M., Singh, A.J. And Handa, D., 2017. Literature Survey On Student's Performance Prediction In Education Using Data Mining Techniques.

Lantz, B., 2015. *Machine Learning With R*. Packt Publishing Ltd.

Márquez, V. C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. 2016. Early Dropout Prediction Using Data Mining: *A Case Study With High School Students. Expert Systems*, 33(1), 107–124.

Meier, Y., Xu, J., Atan, O. And Van Der Schaar, M., 2016. Predicting Grades. *IEEE Transactions On Signal Processing*, *64*(4), Pp. 959-972.

Natek, S. And Zwilling, M., 2014. Student Data Mining Solution–Knowledge Management System Related To Higher Education Institutions. *Expert Systems With Applications*, *41*(14), Pp. 6400-6407.

Nisbet, R., Elder, J. And Miner, G., 2009. *Handbook Of Statistical Analysis And Data Mining Applications*. Academic Press.

Olson, D. L. 2007. Data mining in business services. Service Business, 1(3), 181-193. doi:10.1007/s11628-006-0014-7

Osmanbegović, E. And Suljić, M., 2012. Data Mining Approach For Predicting Student Performance. *Economic Review*, *10*(1).

Oussena, S., 2008. Mining Courses Management Systems, Thames Valley University.

Oyerinde, O.D. And Chia, P.A., 2017. Predicting Students' Academic Performances–A Learning Analytics Approach Using Multiple Linear Regression.

Papamitsiou, Z.K. And Economides, A.A., 2014. Learning Analytics And Educational Data Mining In Practice: A Systematic Literature Review Of Empirical Evidence. *Educational Technology & Society*, *17*(4), Pp. 49-64.

Qatar National Vision 2030, p.11

https://www.psa.gov.qa/en/qnv1/Documents/QNV2030_English_v2.pdf

Qatar University, 2019. Retrieved from: qu.edu.qa/about

QS Asia News Network, Sept. 14, 2018. Higher Education in the Middle East: Challenges and Opportunities for US Universities and Middle Eastern Partners. WOWNEWS Retrieved from: https://wownews.com/higher-education-middle-east-2/

Romero, C. And Ventura, S., 2007. Educational Data Mining: A Survey From 1995 To 2005. *Expert Systems With Applications*, *33*(1), Pp. 135-146.

Romero, C. And 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., Ventura, S. And García, E., 2008. Data Mining In Course Management Systems: Moodle Case Study And Tutorial. *Computers & Education*, *51*(1), Pp. 368-384.

Romero, C., Ventura, S., Espejo, P.G. And Hervás, C., 2008, June. Data Mining Algorithms To Classify Students. In *Educational Data Mining 2008*.

Romero, C., Ventura, S., Pechenizkiy, M. And Baker, R.S. Eds., 2010. *Handbook Of Educational Data Mining*. CRC Press.

Sabourin, J., Kosturko, L., Fitzgerald, C. And Mcquiggan, S., 2015. Student Privacy And Educational Data Mining: Perspectives From Industry. *International Educational Data Mining Society*.

Schendel, R., and T. McCowan. 2016. Expanding higher education systems in low-and-middle-income countries: The challenges of equity and quality. Higher Education 72: 407–411.

Shahiri, Amirah & Husain, Wahidah & Abdul Rashid, Nur'Aini. ,2015. A Review On Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science. 72. 414-422. 10.1016/J.Procs.2015.12.157.

Statsoft, Inc. 2013. Electronic Statistics Textbook. Tulsa, OK: Statsoft. WEB: Http://Www.Statsoft.Com/Textbook/

Stephenson, G. 2016. Can academic freedom make space for minority groups? In University World News, Issue No. 439, http://www.universityworldnews.com/article.php?story=20161129151820905 Accessed 27 December 2016.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. And Abreu, R., 2015. A Comparative Study Of Classification And Regression Algorithms For Modelling Students' Academic Performance. *International Educational Data Mining Society*.

Tagliaferri, L. (2017). An Introduction to Machine Learning. https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning.

Tair, M.M.A. And El-Halees, A.M., 2012. Mining Educational Data To Improve Students' Performance: A Case Study. *International Journal Of Information*, *2*(2).

Thakar, P., 2015. Performance Analysis And Prediction In Educational Data Mining: A Research Travelogue. *Arxiv Preprint Arxiv:1509.05176*.

Tsai, C.F., Tsai, C.T., Hung, C.S. And Hwang, P.S., 2011. Data Mining Techniques For Identifying Students At Risk Of Failing A Computer Proficiency Test Required For Graduation. *Australasian Journal Of Educational Technology*, *27*(3).

Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M. And Rauterberg, M., 2015. Advances In Learning Analytics And Educational Data Mining. *Proc. Of ESANN2015*, Pp. 297-306.

Vandamme, J.P., Meskens, N. And Superby, J.F., 2007. Predicting Academic Performance By Data Mining Methods. *Education Economics*, *15*(4), Pp. 405-419.

Veeramuthu, P., Periyasamy, R. And Sugasini, V., 2014. Analysis of Student Result Using Clustering Techniques.

Walldén, S. And Mäkinen, E., 2014. Educational Data Mining And Problem-Based Learning. *Informatics In Education*, *13*(1), P. 141.

Wang, Y. & Witten, I. H. 1997. Induction Of Model Trees For Predicting Continuous Classes. (Working Paper 96/23). Hamilton, New Zealand: University Of Waikato, Department Of Computer Science.

Yadav, S.K., Bharadwaj, B. And Pal, S., 2012. Data Mining Applications: A Comparative Study For Predicting Student's Performance. *Arxiv Preprint Arxiv:1202.4815*.

Yehuala, M. A. 2015. Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University). *International Journal Of Scientific & Technology Research*, 4(4), 91-94.

Yorke, M., Barnett, G., Evanson, P., Haines, C., Jenkins, D., Knight, P., Scurry, D., Stowell, M. And Woolf, H., 2015. Mining Institutional Datasets To Support Policy Making And Implementation. *Journal Of Higher Education Policy And Management*, *27*(2), Pp. 285-298.

Zafra, A., Romero, C. And Ventura, S., 2011. Multiple Instance Learning For Classifying Students In Learning Management Systems. *Expert Systems With Applications*, *38*(12), Pp. 15020-15031.

Appendices

Appendix A: Ethical Approval Letters from the two universities:

- **Qatar University (QU)**
- **University of East London (UEL)**

جامعة قطر
QATAR UNIVERSITY

Qatar University Institutional Review Board
QU-IRB

October 5, 2016

Dear Ms. Wesam Al Madhoun,

Sub.: **Research Ethics Review Exemption / PhD Project**
Ref.: **Project titled, "Predictive modelling of student academic performance- the case of higher education in Middle East "**

We would like to inform you that your application along with the supporting documents provided for the above proposal, is reviewed and having met all the requirements, has been exempted from the full ethics review.

Please note that any changes/modification or additions to the original submitted protocol should be reported to the committee to seek approval prior to continuation.

Your Research Ethics Approval No. is: **QU-IRB 654-E/16**

Kindly refer to this number in all your future correspondence pertaining to this project.

Best wishes,

جامعة قطر
QATAR UNIVERSITY
Institutional Review Board
(IRB)
Office Of Academic Research

Dr. Khalid Al-Ali
Chairperson, QU-IRB

12th April 2017

Dear Wesam,

Project Title:	**Predictive modelling of student academic performance – the case of higher education in Middle East**
Principal Investigator:	**Professor Allan Brimicombe**
Researcher:	**Wesam T M Al Madhoun**
Reference Number:	**UREC 1617 45**

I am writing to confirm the outcome of your application to the University Research Ethics Committee (UREC), which was considered by UREC on **Wednesday 22 March 2017.**

The decision made by members of the Committee is **Approved**. The Committee's response is based on the protocol described in the application form and supporting documentation. Your study has received ethical approval from the date of this letter.

Should you wish to make any changes in connection with your research project, this must be reported immediately to UREC. A Notification of Amendment form should be submitted for approval, accompanied by any additional or amended documents: http://www.uel.ac.uk/wwwmedia/schools/graduate/documents/Notification-of-Amendment-to-Approved-Ethics-App-150115.doc

Any adverse events that occur in connection with this research project must be reported immediately to UREC.

Approved Research Site

I am pleased to confirm that the approval of the proposed research applies to the following research site.

Research Site	Principal Investigator / Local Collaborator
Qatar University	Professor Allan Brimicombe

Approved Documents

The final list of documents reviewed and approved by the Committee is as follows:

Document	*Version*	*Date*
UREC application form	4.0	12 April 2017
Annexe A - Research ethics approval form for the collection and storage of security-sensitive information	1.0	6 April 2017

Approval is given on the understanding that the UEL Code of Practice in Research is adhered to.

The University will periodically audit a random sample of applications for ethical approval, to ensure that the research study is conducted in compliance with the consent given by the ethics Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of this project.

Yours sincerely,

Fernanda Silva
Administrative Officer for Research Governance
University Research Ethics Committee (UREC)
Email: **researchethics@uel.ac.uk**

Appendix B: Factor Analysis- student who have 6_Term GPA (Admitted in Term 1)

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Age	103	19.000	29.000	21.524	1.737
Earned_Hours	103	12.000	103.000	57.835	22.408
High School %	103	65.710	96.700	84.629	7.298
APL_Accuplacer	103	163.000	440.000	276.235	34.254
APIC_Integ_Core	103	192.000	461.000	328.562	42.944
APLU_Lang_Use	103	35.000	120.000	89.743	13.180
APWS_Writing_Workshop	103	80.000	237.000	178.248	25.258
APRS_Reading_Skills	103	27.000	116.000	82.900	16.327
APSM_Sentence_Meaning	103	36.000	117.000	88.242	14.714
APLG_listening	103	48.000	110.000	75.705	9.277
IELTS	103	5.000	8.500	6.036	0.525
APLA_ Arithmetic	103	18.916	92.000	34.596	9.522
APCL_COLL_Level_MATH	103	20.000	85.000	44.069	9.306
APEA_ELEM_ALGEBRA	103	0.000	117.000	76.272	22.973
Term_1	103	0.000	4.000	2.340	1.107
Term_2	103	0.000	4.000	2.304	1.087
Term_3	103	0.000	4.000	2.105	1.024
Term_4	103	0.000	4.000	2.240	1.003
Term_5	103	0.660	4.000	2.367	0.866

Eigenvalues:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19
Eigen value	4.836	3.907	1.433	1.363	1.204	1.014	0.894	0.756	0.565	0.523	0.468	0.404	0.331	0.320	0.273	0.228	0.209	0.174	0.098
Variability (%)	25.454	20.566	7.542	7.171	6.336	5.337	4.705	3.978	2.976	2.750	2.462	2.128	1.740	1.685	1.435	1.200	1.102	0.918	0.515
Cumulative %	25.454	46.019	53.561	60.733	67.068	72.406	77.111	81.089	84.064	86.815	89.277	91.405	93.145	94.830	96.265	97.464	98.567	99.485	100.000

Factor pattern coefficients:

	F1	F2	F3	F4	F5	F6
Age	-0.078	-0.066	0.007	0.609	-0.241	0.134
Earned_Hours	0.193	-0.407	0.035	0.002	-0.044	-0.031
High School %	0.107	-0.196	0.220	-0.379	-0.461	0.076
APL_Accuplacer	0.043	0.100	-0.432	-0.302	-0.324	-0.251
APIC_Integ_Core	0.397	0.157	0.023	0.160	0.008	0.044
APLU_Lang_Use	0.374	0.161	-0.087	0.023	-0.016	0.139
APWS_Writing_Workshop	0.360	0.150	-0.044	0.182	0.029	0.050
APRS_Reading_Skills	0.341	0.124	0.052	-0.083	0.142	-0.065
APSM_Sentence_Meaning	0.370	0.148	-0.126	-0.037	0.004	-0.136
APLG_listening	0.332	0.183	-0.004	0.094	-0.094	-0.100
IELTS	0.001	-0.182	-0.121	-0.057	0.677	-0.284
APLA_ Arithmetic	0.128	0.025	0.561	0.173	-0.009	-0.442
APCL_COLL_Level_MATH	0.132	-0.028	0.279	-0.138	0.213	0.715
APEA_ELEM_ALGEBRA	0.115	-0.028	0.311	-0.484	0.046	-0.090
Term_1	0.105	-0.361	0.160	0.132	-0.256	-0.139
Term_2	0.186	-0.235	-0.405	-0.064	-0.022	0.170
Term_3	0.135	-0.381	-0.184	0.024	-0.027	0.011
Term_4	0.152	-0.356	-0.072	0.080	0.105	-0.098
Term_5	0.139	-0.381	0.058	0.058	0.092	0.047

Correlations between variables and factors after Varimax rotation:

	D1	D2	D3	D4
Age	-0.094	0.119	0.343	-0.641
Earned_Hours	0.082	0.880	0.134	0.169
High School %	-0.023	0.358	0.073	0.580
APL_Accuplacer	0.152	-0.095	-0.633	0.084
APIC_Integ_Core	0.937	0.029	0.122	0.021
APLU_Lang_Use	0.881	0.014	-0.071	0.090
APWS_Writing_Workshop	0.869	0.029	0.061	-0.054
APRS_Reading_Skills	0.753	0.017	0.020	0.260
APSM_Sentence_Meaning	0.857	0.035	-0.143	0.130
APLG_listening	0.817	-0.070	0.046	0.038
IELTS	-0.130	0.350	-0.124	0.021
APLA_ Arithmetic	0.246	-0.040	0.692	0.179
APCL_COLL_Level_MATH	0.188	0.084	0.236	0.355
APEA_ELEM_ALGEBRA	0.089	0.030	0.081	0.714
Term_1	-0.050	0.718	0.319	0.055
Term_2	0.230	0.642	-0.398	-0.024
Term_3	0.013	0.833	-0.095	0.001
Term_4	0.063	0.786	0.050	0.008
Term_5	0.000	0.794	0.178	0.095

Appendix C: Linear Regression- Students who have 6_Term GPA (Admitted in Term 1)

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Term_6_GPA-Normalized	72	0.000	1.000	0.498	0.257
Earned_Hours	72	12.000	103.000	57.403	23.612
Term_5	72	0.660	4.000	2.338	0.860

Correlation matrix:

	Earned_Hours	Term_5	Term_6_GPA-Normalized
Earned_Hours	**1**	0.708	0.700
Term_5	0.708	**1**	0.998
Term_6_GPA-Normalized	0.700	0.998	**1**

Goodness of fit statistics (Term_6_GPA-Normalized):

Statistic	Training set	Validation set
Observations	72.000	31.000
Sum of weights	72.000	31.000
DF	70.000	29.000
R^2	0.996	0.997
Adjusted R^2	0.996	
MSE	0.000	0.000
RMSE	0.015	0.011
MAPE	2.119	1.918
DW	2.213	
Cp	2.618	
AIC	-599.548	
SBC	-594.995	
PC	0.004	
Press	0.017	
Q^2	0.996	0.000

Type I Sum of Squares analysis (Term_6_GPA-Normalized):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	0	0.000			
Term_5	1	4.671	4.671	19846.602	**< 0.0001**

Type III Sum of Squares analysis (Term_6_GPA-Normalized):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	0	0.000			
Term_5	1	4.671	4.671	19846.602	**< 0.0001**

Model parameters (Term_6_GPA-Normalized):

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Intercept	-0.199	0.005	-37.863	**< 0.0001**	-0.210	-0.189
Earned_Hours	0.000	0.000				
Term_5	0.298	0.002	140.878	**< 0.0001**	0.294	0.302

Figure: Pred(Term_6_GPA-Normalized) / Term_6_GPA-Normalized

Appendix D: Factor Analysis-Students who have 2_term GPA (Admitted in Term 2)

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
IELTS	326	5.000	8.000	5.951	0.349
APLA_ Arithmetic	326	19.713	79.000	36.388	8.920
APCL_COLL_Level_MATH	326	20.000	93.000	48.092	9.761
APEA_ELEM_ALGEBRA	326	0.000	120.000	71.263	21.998
APL_Accuplacer	326	143.000	514.000	289.744	35.903
APIC_Integ_Core	326	128.000	443.000	302.574	39.008
APLU_Lang_Use	326	30.000	116.000	77.478	12.884
APWS_Writing_Workshop	326	56.000	227.000	153.481	22.132
APRS_Reading_Skills	326	26.000	120.000	77.900	13.249
APSM_Sentence_Meaning	326	26.000	118.000	73.200	12.620
APLG_listening	326	33.000	107.000	68.545	8.076
Age	326	18.000	45.000	22.859	4.730
High School %	326	63.630	99.650	84.141	6.504
Earned_Hours	326	0.000	69.000	14.264	10.535
Term1_GPA	326	0.000	4.000	2.016	1.245

Eigenvalues:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
Eigenvalue	4.897	1.769	1.147	1.052	1.029	0.986	0.921	0.670	0.625	0.507	0.422	0.374	0.366	0.189	0.046
Variability (%)	32.650	11.795	7.645	7.015	6.857	6.572	6.142	4.464	4.165	3.378	2.813	2.496	2.439	1.261	0.309
Cumulative %	32.650	44.444	52.090	59.104	65.961	72.533	78.675	83.139	87.304	90.682	93.495	95.990	98.429	99.691	100.000

Appendix E: Data Science algorithms- Students Admitted in Term 2

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Term2_GPA	228	0.000	4.000	1.683	1.457
Earned_Hours	228	0.000	69.000	14.504	10.655
Term1_GPA	228	0.000	4.000	2.064	1.206

Analysis of variance (Term2_GPA):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Model	2	175.282	87.641	64.324	**< 0.0001**
Error	225	306.560	1.362		
Corrected Total	227	481.842			

Type I Sum of Squares analysis (Term2_GPA):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	1	137.820	137.820	101.154	**< 0.0001**
Term1_GPA	1	37.462	37.462	27.495	**< 0.0001**

Type III Sum of Squares analysis (Term2_GPA):

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	1	42.641	42.641	31.297	**< 0.0001**
Term1_GPA	1	37.462	37.462	27.495	**< 0.0001**

Figure: Term2_GPA / Standardized coefficients

Group 1: Students with 2_term GPA

Step 4: Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy:

Correlation: Pearson (n)
Extraction method: Principal factor analysis
Stop conditions: Convergence = 0.0001 / Iterations = 50
Rotation: Varimax (Kaiser normalization)

Correlation Matrix in relation with the dependent variables

Variables	TOEFL	SAT	IC3	ACT	IELTS	Age	PL_Accuplac	arned_Hour	COLL_Level	Term_2_Nor
TOEFL	1	0.000	0.110	-0.054	-0.015	-0.006	0.118	0.046	0.000	0.008
SAT	0.000	1	0.044	-0.016	0.073	0.040	0.041	0.022	0.000	0.027
IC3	0.110	0.044	1	0.087	0.005	0.019	0.292	0.079	0.000	0.061
ACT	-0.054	-0.016	0.087	1	0.183	-0.056	-0.058	-0.061	0.004	0.026
IELTS	-0.015	0.073	0.005	0.183	1	-0.019	-0.066	0.073	0.072	0.092
Age	-0.006	0.040	0.019	-0.056	-0.019	1	0.019	-0.048	-0.017	-0.223
APL_Accuplacer	0.118	0.041	0.292	-0.058	-0.066	0.019	1	0.093	0.016	0.062
Earned_Hours	0.046	0.022	0.079	-0.061	0.073	-0.048	0.093	1	0.075	0.630
APCL_COLL_Level_MAT	0.000	0.000	0.000	0.004	0.072	-0.017	0.016	0.075	1	0.100
Term_2_Normalized	0.008	0.027	0.061	0.026	0.092	-0.223	0.062	0.630	0.100	1
Term_1	-0.014	0.026	0.067	0.017	0.081	-0.195	0.068	0.627	0.142	0.681
High School %	-0.002	0.050	-0.020	0.080	0.106	-0.205	0.027	0.261	0.140	0.451
APLA_Arithmetic	0.001	0.000	0.000	-0.023	-0.004	-0.044	-0.008	0.062	-0.002	0.084
APEA_ELEM_ALGEBRA	0.002	0.000	0.000	-0.017	0.029	-0.079	0.042	0.188	0.159	0.221
APRS_Reading_Skills	-0.002	0.000	0.000	0.014	-0.034	-0.061	0.084	0.060	0.116	0.129
APLG_listening	0.002	0.000	0.000	0.003	-0.020	-0.041	0.060	0.059	0.083	0.104
APLU_Lang_Use	0.000	0.000	0.000	0.014	-0.027	-0.049	0.100	0.061	0.126	0.106
APSM_Sentence_Meanin	-0.001	0.000	0.000	0.010	-0.026	-0.047	0.081	0.062	0.101	0.096
APWS_Writing_Worksho	0.000	0.000	0.000	0.013	-0.027	-0.042	0.010	0.061	0.119	0.095
APIC_Integ_Core	0.000	0.000	0.000	0.012	-0.029	-0.047	0.011	0.064	0.119	0.108

Variables with high KMO Score but with very low correlation with the dependent variable

APIC_Integ_Core	0.993
APLU_Lang_Use	0.916
APWS_Writing_Workshop	0.974
APRS_Reading_Skills	0.875
APSM_Sentence_Meaning	0.929
APLG_listening	0.873

Summary statistics

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Term_2 GPA	990	0.000	1.000	0.587	0.300
High School %	990	60.600	100.000	84.990	6.926
Earned_Hours	990	0.000	132.000	21.564	10.259
Term_1	990	0.000	4.000	2.495	1.122

Correlation matrix

	High School %	Earned_Hours	Term_1	Term_2 GPA
High School %	1	0.277	0.423	0.432
Earned_Hours	0.277	1	0.629	0.646
Term_1	0.423	0.629	1	0.676
Term_2 GPA	0.432	0.646	0.676	1

Regression of variable Term_2 GPA:

Summary of the variable's selection Term_2 GPA:

Nbr. of variables	Variables	Variable IN/OUT	Status	MSE	R^2	Adjusted R^2
1	Term_1	Term_1	IN	0.049	0.457	0.457

2	Earned_Hours / Term_1	Earned_Hours	IN	0.042	0.538	0.537
3	High School % / Earned_Hours / Term_1	High School %	IN	0.040	0.562	0.561

Regression Tree

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 989

node), split, n, deviance, yval

* denotes terminal node

1) root 989 193.407000 1.8902000

2) Term_1_GPA< 2.685 949 108.089400 1.8580170

4) Term_1_GPA< 2.033237 60 75.374570 1.4815000

8) Earned_Hours< 7.5 15 3.433333 0.2666667 *

9) Earned_Hours>=7.5 45 42.424830 1.8864440

18) Term_1>=1.125 28 22.657840 1.5464290

36) Earned_Hours< 19.5 18 16.035490 1.2605560 *

37) Earned_Hours>=19.5 10 2.503490 2.0610000 *

19) Term_1< 1.125 17 11.198190 2.4464710 *

5) Term_1_GPA>=2.033237 889 23.634860 1.8834290

10) Term_1_GPA>=2.205 16 19.583140 1.4531250 *

11) Term_1_GPA< 2.205 873 1.034841 1.8913150 *

3) Term_1_GPA>=2.685 40 61.014340 2.6537500

6) Earned_Hours< 15 9 14.892400 0.8366667 *

7) Earned_Hours>=15 31 7.778548 3.1812900

14) High School %< 89.015 9 1.970756 2.6177780 *

15) High School %>=89.015 22 1.780727 3.4118180 *

Regression tree:

rpart(formula = Term_2_GPA ~ ., data = crs$dataset[crs$train,

c(crs$input, crs$target)], method = "anova", model = TRUE,

parms = list(split = "information"), control = rpart.control(usesurrogate = 0,

maxsurrogate = 0))

Variables actually used in tree construction:

[1] Earned_Hours High School % Term_1 Term_1_GPA

Root node error: 193.41/989 = 0.19556

Random Forest

Summary statistics (Training/ Quantitative)

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Term_2 GPA	1411	0.000	1.000	0.588	0.302
High School %	1411	60.600	100.000	85.052	6.997
Earned_Hours	1411	0.000	132.000	21.644	10.378

Summary statistics (Prediction set/ Quantitative)

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
High School %	499	60.600	98.950	83.891	6.953
Earned_Hours	499	0.000	132.000	21.160	10.471
Term_1	499	0.000	4.000	2.483	1.042

The Mean Square Error which measures the prediction error is indicated below.

MSE:	0.025

The error rate is explained by the curve below. As we see, the error decreases when the growth of the number of trees.

MSE error evolution

Variable importance

Variables	Mean increase error
High School %	6.646
Earned_Hours	77.891
Term_1	35.765

Group 3: Students with 4_term GPA

Correlation matrix between the dependent variable and extracted independent variables

	Earned_Hours	Term 3_GPA	Term 4_GPA_Normalized
Earned_Hours	1	0.675	0.648
Term 3_GPA	0.675	1	0.611
Term 4_GPA_Normalized	0.648	0.611	1

Type I Sum of Squares analysis (Term 4_GPA)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	1	5.839	5.839	144.074	< 0.0001
Term 3_GPA	1	0.771	0.771	19.033	< 0.0001

Type III Sum of Squares analysis (Term_4 GPA)

Source	DF	Sum of squares	Mean squares	F	Pr > F
Earned_Hours	1	1.419	1.419	35.016	< 0.0001
Term 3_GPA	1	0.771	0.771	19.033	< 0.0001

Term 4_GPA Standardized Coefficient and the second graph shows predicted Term_4 GPA

Regression Trees

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 184

node), split, n, deviance, yval

* denotes terminal node

1) root 184 208.54640 2.0907070

2) Earned_Hours< 33.5 92 75.44122 1.5167390

4) Earned_Hours< 17 21 10.91247 0.6966667 *

5) Earned_Hours>=17 71 46.22866 1.7592960

10) Term.3_GPA>=2.68 10 8.49321 1.0370000 *

11) Term.3_GPA< 2.68 61 31.66308 1.8777050

22) Term.3_GPA< 1.64 32 13.21619 1.6693750 *

23) Term.3_GPA>=1.64 29 15.52553 2.1075860 *

3) Earned_Hours>=33.5 92 72.48849 2.6646740

6) Term.3_GPA< 2.955 51 31.10570 2.3031370 *

7) Term.3_GPA>=2.955 41 26.42461 3.1143900

14) Term.3_GPA< 3.73 31 21.46184 2.9429030 *

15) Term.3_GPA>=3.73 10 1.22504 3.6460000 *

Regression tree:

```
rpart(formula = Term.4_GPA ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "anova", model = TRUE,
    parms = list(split = "information"), control = rpart.control(minbucket = 9,
        usesurrogate = 0, maxsurrogate = 0))
Variables actually used in tree construction:
[1] Earned_Hours Term.3_GPA
```

Random Forest

```
randomForest(formula = Term.4_GPA ~ .,
             data = crs$dataset[crs$train, c(crs$input, crs$target)],
             ntree = 500, mtry = 1, importance = TRUE, replace = FALSE, na.action = randomForest::na.roughfix)
              Type of random forest: regression
                    Number of trees: 500
No. of variables tried at each split: 1
          Mean of squared residuals: 0.77035    % Var explained: 32.03
Variable Importance
         %IncMSE IncNodePurity
Earned_Hours   41.66        51.48
Term.3_GPA     23.76        45.63
```

Group 4: Students with 6_term GPA

Summary statistics

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
Term6_GPA	168	0.500	4.000	2.293	0.755

Shapiro-Wilk test outputs

W	0.986
p-value (Two-tailed)	0.080
alpha	0.05

Regression Tree

Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 118

node), split, n, deviance, yval

* denotes terminal node

1) root 118 66.528490 2.299746

2) Term5_GPA< 1.79 36 13.225830 1.708611

4) Term3_GPA< 0.89 7 1.255743 1.337143 *

5) Term3_GPA>=0.89 29 10.771010 1.798276 *

3) Term5_GPA>=1.79 82 35.199960 2.559268

6) Term5_GPA< 3.05 63 20.655640 2.388413

12) Term1_GPA< 2.55 42 11.376300 2.243095 *

13) Term1_GPA>=2.55 21 6.618581 2.679048 *

7) Term5_GPA>=3.05 19 6.607263 3.125789 *

Regression tree:

rpart(formula = Term6_GPA ~ ., data = crs$dataset[crs$train,

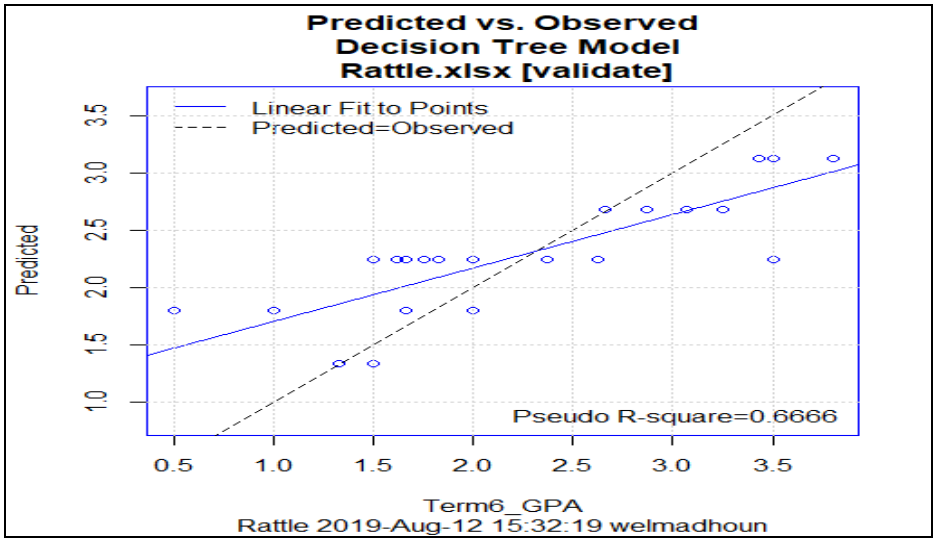
c(crs$input, crs$target)], method = "anova", model = TRUE,

parms = list(split = "information"), control = rpart.control(maxdepth = 3,

cp = 0.009, usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:

[1] Term1_GPA Term3_GPA Term5_GPA



Predicted vs. observed values_ Regression Tree- students with 6_term GPA and admitted in term 2

Random Forest

```
Summary of the Random Forest Model

Number of observations used to build the model: 135

randomForest(formula = Term6_GPA ~ .,
             data = crs$dataset[crs$train, c(crs$input, crs$target)],
             ntree = 550, mtry = 1, importance = TRUE, replace = FALSE, na.action = na.omit)
              Type of random forest: regression
                    Number of trees: 550
No. of variables tried at each split: 1
          Mean of squared residuals: 0.338156
                    % Var explained: 36.77
```