Predicting Dropouts among a Homogeneous Population using a Data Mining Approach

توقع نسبة الطلاب المنسحبين من خلال تطبيق التنقيب عن البيانات على مجموعه متجانسة من الطلاب

by

GHAZALA BILQUISE

Dissertation submitted in fulfilment

of the requirements for the degree of

MSc INFORMATICS
(KNOWLEDGE AND DATA MANAGEMENT)

at

The British University in Dubai

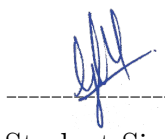
March 2019

Declaration

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions. I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate. I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.



Student Signature

Copyrights and User Information

Abstract

Student retention is one the biggest challenges facing academic institutions worldwide. Failure to retain students not only affects the student in a negative way but also hinders institutional quality and reputation. While there are several theoretical perspectives of retention, which study the factors that cause students to drop out, more recent studies rely on a data mining and machine-learning approach to explore the problem of retention.

In this research, we present a novel data mining approach to predict retention among a homogeneous group of students, with similar social and cultural background, at an academic institution based in the UAE. Our model successfully identifies dropouts at an early stage. It provides an early warning system that enables the institution to promptly intervene with assertive measures. Moreover, our model also effectively determines the top predictive variables of retention.

Several researchers study retention by focusing on student persistence from one term to another while our study builds a predictive model to study retention until graduation. Moreover, other works use additional student data for predictions, thereby reducing the dataset size, which is counter productive to data mining. Our research relies solely on pre-college and college performance data available in the institutional database.

Our research reveals that the Gradient Boosted Trees is a robust algorithm that predicts dropouts with an accuracy of 79.31% and AUC of 88.4% using only pre-enrollment data. High School Average and High School stream of study are observed to be the top predictive variables of on-time graduation when a student joins college. Our study also reveals that ensemble machine-learning algorithms are more reliable and outperform standard algorithms.

نبذة مختصرة

يعد الاحتفاظ بالطلاب أحد أكبر التحديات التي تواجه المؤسسات الأكاديمية في جميع أنحاء العالم. الفشل في الاحتفاظ بالطلاب لا يؤثر فقط على الطالب بطريقة سلبية ولكنه يعوق أيضًا الجودة والسمعة المؤسسية. في حين أن هناك العديد من الدراسات النظرية للاحتفاظ بالطلاب، والتي تدرس العوامل التي تسبب انسحاب الطلاب ، إلا أن المزيد من الدراسات الحديثة تعتمد على نهج التنقيب عن البيانات (Data mining) والتعلم الآلي (Machine Learning)لاستكشاف مشكلة الاستبقاء.

في هذا البحث ، نقدم طريقة جديدة في التنقيب عن البيانات للتنبؤ ببقاء الطلاب بين مجموعة متجانسة من الطلاب ، مع خلفية اجتماعية وثقافية متماثلة ، في مؤسسة أكاديمية مقرها الإمارات العربية المتحدة. يتنبأ نموذجنا بالمنسحبين في مرحلة مبكرة، بنجاح. إنه يوفر نظام إنذار مبكر يمكّن المؤسسة من التدخل الفوري بتدابير حازمة. علاوة على ذلك ، فإن نموذجنا يحدد بفعالية أهم المتغيرات التنبؤية لبقاء الطلاب.

يدرس العديد من الباحثين الاستبقاء من خلال التركيز على بقاء الطالب من فصل إلى آخر بينما تقوم دراستنا ببناء نموذج تنبؤي لدراسة الاستبقاء حتى التخرج. علاوة على ذلك ، تستخدم أعمال أخرى بيانات الطلاب الإضافية للتنبؤات ، مما يقلل من حجم مجموعة البيانات ، مما يؤدي إلى نتائج عكسية في التنقيب عن البيانات. يعتمد بحثنا فقط على بيانات أداء ما قبل الكلية و الاداء داخل الكلية، و المتوفرة في قاعدة البيانات المؤسسية.

يكشف بحثنا أن أشجار التدرج المعزز (Gradient Boosted Trees) هي خوارزمية قوية تتنبأ بالمنسحبين بدقة تبلغ 79.31٪ و AUC بنسبة 88.4٪ باستخدام بيانات ما قبل التسجيل فقط. يُلاحظ أن معلومات معدل الدراسة في المدرسة الثانوية بالاضافة الى نوع الدراسة في المدارس الثانوية - والتي تتوفر عند انضمام الطالب للكلية- هما أهم المتغيرات التنبؤية لتخرج الطالب في الوقت المحدد. تكشف دراستنا أيضًا أن خوارزميات التعلم الآلي المكونة من عدة أجزاء (Ensemble Machine-Learning) هي أكثر موثوقية وتتفوق على الخوارزميات التقليدية.

Acknowledgements

All praises are to Allah for bestowing me with the strength, knowledge, ability and opportunity to undertake this research and to persevere and complete it satisfactorily. Without His blessings, this achievement would not have been possible.

First of all, I would like to express my sincere gratitude to my supervisor, Professor Sherief Abdallah, for his time and guidance throughout the whole dissertation project. I have learned a lot from him on the topic of data science and machine learning. I would also like to thank all my teachers in the masters program.

A special thanks goes to my esteemed colleague Dr. Thaeer Kobbaey for his motivation, encouragement and guidance. I appreciate his sincere and selfless support in reviewing and polishing sections of my writing and for troubleshooting Latex issues.

I would like to pay the warmest tribute to my beloved family whose patience and support made it possible for me to persist in my studies.

Last but not the least, I would also like to thank my colleague and dear friend Aisha Ghazal without whom I would not have embarked on this academic journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The need for economic development, innovation and technological advancement are fuelling the demands for a highly skilled and qualified workforce. In today's knowledge society, education is the key to creativity and innovation, which are essential elements of progress. Educational attainment is increasing in importance now more than ever as nations have come to realize that only educated people can innovate new technologies. Moreover, earning a college degree provides individuals with opportunities for career growth and lowers unemployment rates thereby boosting the local economy (Ma et al., 2016).

Despite the enormous socio-economic benefits of earning a college degree, navigating students from enrollment to graduation is a challenging task for Higher Educational Institutions (HEI). Nearly 30% of students leave college without earning any credential (ACT, 2018). The high dropout rate of students is prevalent worldwide (Chalaris et al., 2015) with the United Arab Emirates (UAE) being no exception (GulfNews, 2017).

Academic institutions apply various preventative measures to retain students by creating an environment that addresses retention challenges such as setting up student clubs, counselling, providing financial aid and academic support. However, these measures have shown little or no improvement in retention rates (Demetriou and Schmitz-Sciborski, 2011; Seidman, 2005; Yu et al., 2010). One of the reasons for this is that multiple factors play

a role in students not completing their education on time. These include both personal factors and institutional challenges such as financial considerations, psychological factors, inability to cope with academic requirements and more (Tinto, 1975).

Moreover, the reasons for dropping out are often inter-related. Therefore, a personalized approach is required by firstly identifying students who are at risk of dropping out, and then providing targeted remedial support that can change the course of a students' academic journey.

Studies show that students resolution to drop out of college mostly occurs in the freshman year as they struggle to cope with the challenges of an academic environment and transition from high school (Delen, 2011; Hoffait and Schyns, 2017; Tinto, 1975). Therefore, intervention strategies within the first year of education can improve retention rates by up to 50% (Levitz et al., 1999). An early identification of students at risk of premature departure will enable the institution to target their resources to benefit students who need it the most.

While retention remains a challenge for HEIs, a promising avenue to address this challenge is the use of data mining and machine-learning techniques. Data mining is the process of discovering patterns from large datasets by identifying relationships that would otherwise be impossible (Romero and Ventura, 2013). It integrates the disciplines of computer science and statistics to make predictions by discovering previously unknown trends in data (Larose and Larose, 2014). Data mining techniques are applied extensively in the field of business to study customer behaviour and in particular customer churn (Vafeiadis et al., 2015).

Educational Data Mining (EDM), a subdomain of data mining, is an emerging discipline that leverages the power of data obtained from educational settings to analyze and study student behaviour (Baker and Yacef, 2009; Jacob et al., 2015; Romero and Ventura, 2013). The predictive models generated by machine-learning algorithms enable HEIs to take strategic actions for future improvements.

This research aims to develop a model to predict student retention using an EDM approach and to identify the top factors that contribute to students dropping out of a college based in the UAE. With this insight, the college can set up early intervention strategies to assist students in completing their degree on time.

1.0.1 Definition of Terms

Some terms associated with retention are often used interchangeable. Therefore, it is essential to define and differentiate between these terms as used in our research.

Student Retention is defined as the ability of an academic institution to successfully retain a student until graduation.

Persistence denotes the number of students who re-enroll the following term and hence persist and continue their studies. Continued student persistence until graduation would lead to retention.

Attrition rate, also known as dropout rate, is the percentage of students who discontinue their studies. In our study we refer to a dropout as a student who has left his studies without attaining a degree.

Persistence and retention are often used interchangeably by several studies. However, we clearly distinguish between these two terms since our research is about retention. The study of persistence requires data from each term and the following term only, while the study of retention requires longitudinal data of the same student across multiple terms until graduation.

A student may discontinue his/her studies on his own accord or be withdrawn by the institution due to poor academic performance or behavioural issues. However, a withdrawn student may return to continue studies in a later semester and hence is considered retained in our research if he graduates on time. While in a study of persistence, such

a student would be considered a dropout. Moreover, if a student re-enrolls the following semester and changes his program of study, this student would be considered as persisted in other studies, while in our study, we consider the student as a dropout in one program and new enrollment in another.

Although many studies use data mining to predict student persistence from one term to another, only a few papers study retention until graduation. This paper studies student retention with the aim to predict students who are not likely to graduate in a bachelor degree program in a UAE based HEI.

1.1 Research Motivations

Student retention and on-time graduation is a critical issue among Higher Education Intuitions (HEIs) worldwide. Around 29% of students who enrolled in a four year degree program in US colleges in the year 2016 did not re-enroll the following year (NSCResearchCenter, 2019). Moreover, the National Center for Education Statistics reported that in 2016, only 60% of students enrolled in a bachelor degree program graduated on time (NCES, 2018). In Belgian Universities average graduation rate is nearly 73% (Hoffait and Schyns, 2017) while in Greece only 47% students graduated on time in the year 2013 (Chalaris et al., 2015).

Similar retention figures have been reported in 2014 within UAE academic institutions with graduation rates ranging from 79% to 84% (UAEU, 2015; ZU, 2014). In 2017, the dropout rate in federal institutions in the UAE was observed to be higher than the average global dropout rates (GulfNews, 2017). Hence, improving retention is crucial in UAE as educational institutions prioritize efforts to achieve this goal.

Student attrition and failure to graduate has adverse effects on both the student as well as the academic institution. Poor retention not only affects the student's career and future prospects but also has a negative impact on the intuitions reputation. Moreover,

retention rate is an essential factor for accreditation, since it is an indicator of institutional quality and performance (Mayra and Mauricio, 2018).

Retaining students is crucial for the financial well-being of an institution. Reducing dropouts by as much as 1% can decrease the financial strain on the institution (Levitz et al., 1999). Hence, HEIs strive to devise timely intervention strategies that promote student success and persistence in their studies.

The issue of retention is essential not only for the student but also for the institution and society at large. There are three main reasons why this topic is of interest to us.

Firstly, improving retention rates contributes to the institutional effectiveness, as it is one of the KPIs (Key Performance Indicator) at the institution of this research. Attrition is a major concern as nearly 30% students dropout without earning a degree in the year 2011. Moreover, current strategies for retention are not satisfactory as they are not focused on potential dropouts. The institution could benefit from this research as a prediction of dropouts will enable them to develop targeted strategies to steer students to graduation. Moreover, the institution will be able to employ their resources and retention efforts more efficiently.

Secondly, the value that a college degree adds to a student's career and personal growth and thereby to the growth of UAE is another motivating factor of this research. A degree provides students with a gateway to social security and economic prosperity leading to job opportunities and a better quality of life. Moreover, education empowers students to think critically, make decisions, evaluate, and utilize information to lead a healthier and happier life (Mirowsky and Ross, 2005).

Lastly, the subject of retention lacks critical investigation in the UAE, to which our research can contribute. To the best of our knowledge, the subject of retention has not been addressed using a data mining approach in the UAE. Moreover, the results of studies on retention based in other countries cannot be generalized to our institution because of the difference in cultural, social and economic environments.

1.2 Research Focus

This research is based on a federal HEI in Dubai that offers six undergraduate programs of study. The students of our academic institution form a homogeneous group belonging to the same nationality, culture and heritage. The education of all students at the institution is funded either federally or through a local sponsor. Nearly 6000 students enroll each term into the various programs with an average of 1300 new enrollments each year. Among this only 70% have completed their graduation on time in the academic year 2011 and 2012.

Several studies that predict retention have differentiated students based on culture, race, ethnicity and more (Kovacic, 2012; Raju and Schumacker, 2015; Tamhane et al., 2014). However, our study is based on a homogeneous population with similar ethnic and social background. Therefore, the factors that influence students to drop out in other cultures do not apply to our environment.

Moreover, while we believe some factors like learning styles, financial status and family education may influence a students decision to drop out, these variables were not available in the institutional database at the time of this research. Acquiring such data would require conducting surveys which is a time-consuming and expensive process. Furthermore, it would also lead to a reduction in the size of our dataset, which in turn would be counterproductive to a data mining approach.

We also believe that regardless of the social, personal and psychological factors that may lead to dropout, these factors will eventually be reflected in the students' academic performance. Hence, our research focuses solely on performance data to predict retention, thereby providing a quick and practical solution to the institution.

Several studies apply machine-learning to predict persistence from one term to another (Guarín et al., 2015; Hoffait and Schyns, 2017). As explained in section 1.0.1, in this

research we clearly distinguish between the terms persistence and retention. Our research aims at predicting student dropout who fail to earn a degree.

The purpose of this research is to apply an EDM approach to predict dropouts among a homogeneous undergraduate population using data that is currently available in the institutional database. Our research relies on student demographics and performance in high school, standardized tests as well as college performance to predict dropouts. In addition, we seek to determine the earliest stage when an effective prediction is possible and identify the top predictive factors of retention at each stage.

1.2.1 Research Objectives and Research Questions

This research employs data mining methods and machine learning algorithms to predict the first year undergraduate students who are likely drop out before graduation. We develop a model, which predicts potential dropouts and allows decision makers to intervene and guide students back on track at an early stage. We base our predictions on enrollment data and student performance data prior to joining college and during the first year of college. We also aim to identify the top predictors of retention. We focus on three main research questions in this study.

Research Question 1: Can machine-learning algorithms effectively predict retention/dropouts among a homogeneous population of students?

Research Question 2: How early can we predict potential dropouts using machine learning?

Research Question 3: Which attributes are the top predictors of retention?

1.3 Research Contributions

A large number of studies that have undertaken the task of investigating student success in undergraduate studies by predicting student performance (Zollanvari, 2017; Miguéis,

2018; Natek, 2014; Huang, 2013; Asif 2017) or ability to progress from one term to another (Kovacic, 2012; Yu, 2010; Yukselturk, 2014; Zhang, 2010; Rubiano, 2015; Guarin, 2015; Hoffait, 2017). Only a few studies have focused on retention until graduation in higher education (Auluck, 2016; Raju, 2015; Perez, 2018; Bayer, 2012; Dekker, 2009). Moreover, to the best of our knowledge, no such study has been conducted till date in UAE or the middle east.

The uniqueness of this research differentiates it from the vast majority of studies carried out in other parts of the world. Almost all these studies focus on retention efforts in their local environment; thus making it hard to map those studies to the UAE. The factors that influence students to drop out in those cultures do not apply to our environment.

Moreover, previous studies are based on private academic institutions that intake students from different cultures and backgrounds, while our research examines UAE nationals in a local institution. Thus, our students form a homogenous group, unlike other countries where students belong to different cultural and ethnic backgrounds. Factors like ethnicity, financial aid, commuting distance, living arrangements and more, do not play an important role in retention in our scenario. Hence, we believe that our study is novel as it investigates retention in a unique educational setting.

Our research provides insights in understanding retention and the predictive factors of retention in this region. It will not only assist the decision makers in our institution to improve retention rates and focus their intervention strategies but also provide a foundation for other researchers who intend to investigate the topic further. In addition to contributing to the body of literature on retention, our research also contributes to the broader body of studies on prediction and classification using an EDM approach.

The current research goes beyond the previous work in retention in several ways. We use a larger dataset than most studies and focus on student retention across multiple disciplines rather than just one specific discipline. We rely solely on institutional data to

develop a more practical and reusable solution that can be applied without the need to collect additional data. We also apply balancing techniques to obtain more reliable and unbiased results while many other studies have failed to do so.

1.4 Research Methodology

The likelihood of obtaining reliable and accurate results increases with the use of a structured approach. In this research, we follow the CRISP-DM (cross-industry process for data mining) approach (Shearer, 2000) (Shearer, 2000), which provides an organized framework for performing data mining tasks. The CRISP-DM methodology consists of six phases outlined below:

1. **Business Understanding:** We assess the needs of the business and determine the goals of data mining and create a plan for our research.
2. **Data Understanding:** We acquire the data and explore it to identify quality issues. A strategy to deal with these issues is set up at in this phase.
3. **Data Preparation:** The data quality issues are resolved and data is pre-processed and prepared for generating the models.
4. **Modeling:** We select the machine-learning algorithms and apply them to the processed dataset to generate predictive models. We test the quality and validity of the models by using a training and testing set data and report the results.
5. **Evaluation:** we compare the results of the various algorithms discuss the findings.
6. **Deployment:** We study the strategies of deploying the model in the institution.

Three software tools are used in our research. Firstly, we use Microsoft Excel 2016 to gather, integrate and filter enrollment records. Secondly, we use Microsoft Access 2016 to

perform the enrollment and graduation records integration and initial data preparation tasks. Lastly, we use Rapid Miner 8.1 to perform additional data preparation tasks and to apply machine-learning algorithms, optimize parameters, balance the dataset and evaluating the model performance. All the aforementioned software runs on a computer with the following specifications: Windows 10 Enterprise; 64-bit Operating System; 16GB RAM; an Intel Core i7 7700HQ CPU @ 2.80GHz processor.

1.5 Organization of the Study

This research is organized as follows; Chapter 2 provides the literature review. It begins with examining the theoretical perspectives on retention and continues by presenting the related work on retention using a data mining approach. It explores the factors used for predicting retention and the techniques applied. Chapter 3 presents our methodology and describes the tasks performed in each phase of the CRISP-DM approach. Chapter 4 gives the results and the discussion of the findings related to the research questions. Finally, Chapter 5 concludes the research and provides the future work.

Chapter 2

Literature Review

Student retention, also referred to as student mortality in earlier studies (Berger et al., 2005), is a complex and multi-dimensional (Astin, 1985; Bean, 1980; Tinto, 1975) phenomenon that has been investigated extensively for the last four decades. While studies in the past have focused on building theoretical frameworks to determine factors that affect retention, studies that are more recent are leveraging the power of data mining to predict retention using data generated from educational settings.

This chapter presents a review of the literature of previous work that has established the groundwork on retention as well as the empirical studies that have investigated retention using data mining and machine learning approach.

2.1 Theoretical perspectives on student retention

Several factors affecting graduation were identified in studies that have proposed a framework for retention (Astin, 1985; Bean, 1980; Tinto, 1975). Although our research leverages the findings of previous theoretical studies, the purpose of this paper is not to develop a new theory but rather to harness the power of data mining techniques to provide alternative ways of addressing the issue of retention in HEIs.

A large number of theories exist on student survival in Higher Education. However, the theoretical framework proposed by Tinto (1975) is one of the most widely accepted and influential model. Tinto ascribed the reason for dropout as the failure of a student to integrate within the institution both academically as well as socially. Moreover, the study argued that the students personal and pre-college characteristics contribute to institutional and academic commitment and thus influence graduation rates.

Bean (1980) compared attrition to employee turnover in a business organization. The study stated that student's satisfaction level with the institution and external factors determines retention rate. Factors that lead to attrition were identified as student background information, academic achievements, student perception of studies and its relevance. The leading factors were identified as the students' academic performance and socioeconomic variables.

Astin (1985) Astins (1985) theory of involvement stated that student involvement in the academic institution leads to reduced attrition rates. The theory further proposed that a combination of student characteristics and prior experiences along with the institutional experiences lead to student success and graduation.

All theories on retention have attributed the phenomenon to several factors including student demographics, academic performance, social integration, economic and financial status as well as the psychological state. Both Tinto (1975) and Bean (1980) argued that there is a strong relationship between student characteristic before enrollment and retention. These characteristics affect how well a student would integrate with the institution. Pre-enrollment characteristic includes high school performance, family and financial background and more.

Socio-economic and financial factors are also identified as a cause of attrition. Tinto (1975) also asserted that dropout rates at public universities are significantly higher due to the lack of financial commitment. Thus, the study implied that financial commitment induces a student to persist and continue his/her education until the student either completes or is dismissed due to academic failure.

Our study is based on a government college where the students' education is either funded by the government or by a sponsor. While the lack of financial commitment may be a cause for increased dropout rates at other institutions, it cannot be considered as a predictive factor in our research since it applies to all students equally.

Despite the various factors identified, all theories agree that both pre-college and college factors attribute to student success and the decision to persist. While the student demographic characteristics and prior knowledge determine how well the student is prepared for academic life, the academic performance and experiences of the student while in college determine if the student will persist.

Our study relies on demographic and performance data to predict dropouts. Other factors such as psychological, financial, social and economic, cited in the literature would have no doubt been valuable. However, these were not readily available at the time of this study. Moreover, academic performance variables are the most important variables of retention as the other factors may eventually be reflected in the academic performance (Larsen et al., 2012).

2.2 EDM Approach to study Retention

Statistical methods and logistic regression are a popular choice among researchers to study the factors leading to retention (Araque et al., 2009; Gershenfeld et al., 2016). However, recent empirical studies have applied EDM approaches to tackle the phenomenon by employing various machine-learning techniques. EDM is a field of study that applies machine learning to identify patterns and gain insights from large volumes of educational data that would otherwise be very challenging due to their massive size (Romero and Ventura, 2010).

Over the past two decades, there has been an increasing interest in the use of EDM to gain insights for educational institutions. Many papers have surveyed the use of EDM

to enhance teaching and learning practices in academic institutions. Although EDM tasks are used in a variety of applications, the most common type of study is the use of classification to predict student performance (Peña-Ayala, 2014; Romero and Ventura, 2007, 2010).

Romero and Ventura (2007) captured the various studies in the field of EDM for one decade from 1995-2005. Baker and Yacef (2009) further investigated EDM studies covering research in this field conducted until 2009. Dutt et al. (2017) provided a review of 3 decades on clustering algorithms used in EDM. Overall, it evident that data mining has been successfully applied in the field of education to provide insights to develop better and improved educational practices.

Peña-Ayala (2014) surveyed 222 papers on EDM, published in the period between 2010 to 2013. More recently, Bakhshinategh et al. (2018) provided a systematic review of the studies in EDM over the past ten years. They categorized research papers based on tasks and applications in EDM and showed that predicting student performance is one of the major applications in this field. All studies showed the growth and potential of using EDM techniques and the insights it provides to decision makers.

In this section, we review the research contributions in the domain of retention and student success using classification techniques. We present our literature review based on five aspects – EDM studies on retention and persistence, Factors used in predicting retention, early detection for early intervention and balancing techniques in these studies.

2.2.1 Persitence and Retention

Several research papers have studied attrition and student success by mainly focusing on persistence rather than retention. Nevertheless, all studies reported their research objective as improving retention rates with an aim of early intervention.

2.2.1.1 Studies on Persistence

Machine learning techniques have been used to predict attrition by investigating failure rates in cornerstone courses. Costa et al. (2017) employed various data mining algorithms to predict student persistence in a first year programming course by using weekly performance grades and demographic data. Aguiar et al. (2014) studied dropouts in engineering by predicting the success of students in an introductory engineering course. Huang and Fang (2013) predicted the final performance of students in a high-impact engineering course using cumulative GPA, grades attained in pre-requisite courses as well as course work assessment scores.

The aforementioned studies are based on the assumption that success in fundamental courses would eventually lead to retention in the corresponding programs of study. However, our research focuses on retention across several disciplines; therefore, we do not consider the outcome of a single course.

Several studies have investigated retention in a degree program, high school or online program by exploring student persistence on a term by term basis with re-enrollment in subsequent terms. Jayaprakash et al. (2014) , Kovacic (2012) and Yukselturk et al. (2014) applied machine-learning algorithms to study persistence in online studies. Delen (2011) and Thammasiri et al. (2014) use machine learning techniques to predict freshmen students' persistence in until the next term.

Márquez-Vera et al. (2016) used several data mining and a genetic programming algorithms to predict high school students who are at risk of dropping out at an early stage. The study collected data in six stages of progression in studies, starting with secondary school performance, pre-enrollment data and gradually added more data. The study achieved a very high accuracy of 99.8% using their genetic algorithm, which outperformed all other algorithms. Given that the study applied balancing techniques, we believe that the high-performance figures may be a sign of overfitting.

Márquez-Vera et al. (2013) applied decision tree algorithms and rule induction algorithms to predict dropouts in the first year of high school using both institutional data as well as data collected from surveys. The research used various factors to study retention in high school, which included social status, family background, psychological profile as well as academic performance. However, the top predictive factors for retention were the performance in Physics, Humanities, Math and English. This finding further reinforces our decision to use academic performance data rather than compromising the size of our dataset by collecting other data.

Our study differs from the aforementioned studies on persistence, as they do not consider graduation as an outcome. They measure dropout as failure to re-enroll in the following term. Moreover, research in persistence does not utilize longitudinal student data across many terms until graduation. The outcome of the prediction is based on a single term or year of data. In our research, we take into consideration that a student who has dropped out for one term may eventually return and graduate on time and hence is not considered as a dropout.

2.2.1.2 Studies on Retention

Research on retention using an EDM approach requires longitudinal data from multiple terms of studies leading to graduation. Aguiar et al. (2015) and Lakkaraju et al. (2015) studied retention in high school using longitudinal data from 6th grade to 12th grade. Both studies utilized institutional data with a relatively large dataset of 11,000 and 200,000 respectively and applied predictive analytics to determine dropouts in high school.

A popular study by Dekker et al. (2009) predicted the successful graduation of first-year students from an electrical engineering diploma program. The study utilized 648 student records, consisting of pre-university data and university performance data, and achieved an accuracy of 80%. One of the main cause of misclassification reported by the authors was the handling of missing course grade values by replacing it with zeros. Our

research avoids this problem by imputing missing values of important features using the k-NN algorithm.

Perez et al. (2018) investigated dropouts who did not complete their degree in a System Engineering undergraduate program. The study was based on 802 student records and included attributes such as admission data, course grades, and financial aid per term. The dataset was reported to be balanced and the study achieved an AUC score of 94%.

Aulck et al. (2016) used eight years of enrollment data comprising of over 69,000 heterogeneous students to predict graduation at a University based in the United States. The dataset consisted of demographic data, such as ethnicity, gender and residency information. The study also used external assessment data, such as SAT and ACT scores and academic achievement scores in all courses to predict dropouts. The study achieved a maximum AUC score of 72.9% using Logistic Regression algorithm.

Djulovic and Li (2013) investigated retention of university students using Decision Trees, Naïve Bayes, Neural Networks and Induction Rule algorithms. The study used a dataset of 7800 enrollment records, ranging from the academic year 2006 to 2012, to predict dropouts. The dataset consisted of demographic data, financial data, SAT scores and three semesters performance data. The highest accuracy of nearly 86% was achieved by the Decision Tree and Induction Rule algorithms. However, the study did not apply balancing techniques to the imbalanced dataset, which consisted of more retained students than dropouts. This leads us to believe that the reported performance results may be biased.

Raju and Schumacker (2015) investigated the factors of retention until graduation using Logistic Regression and Neural Networks algorithms. The study utilized 7,293 records, comprising of pre-college data and the first term of college data, to predict graduation. Pre-college data included demographic information, external assessment scores, high school data, distance from home and work status, while college data included first

semester GPA and earned hours. Missing data was handled by deleting the record that contained missing values. The study achieved an AUC of 77.7% using a Neural Network.

Our research shares similarities with the study of Raju and Schumacker (2015). However, our study also stands apart with the use of ensemble data mining algorithms, optimization of machine-learning algorithms, imputing missing values and using SMOTE to balance the dataset. We achieve results of upto 92% AUC using Gradient Boosted Trees.

The study by Miguéis et al. (2018) and Asif et al. (2017) did not consider dropouts; instead the purpose was to classify undergraduate students' final performance. Asif et al. (2017) used classification and clustering using high school performance data, and academic achievement of all courses, taught in a four-year bachelor degree program, to classify students into five performance groups. The study did not investigate dropouts and failures and is based on a small dataset of 210 records only.

Miguéis et al. (2018) used 2,549 enrollment records of an Engineering school in Europe to classify students into different final year performance groups. The study aimed to provide early and relevant intervention in the first year. Unlike Asif et al. (2017), the study by Migues utilized demographic data, socio-economic status, high school performance, external assessment data as well as the average grade in the first and second term of university.

While both Asif et al. (2017) and Miguéis et al. (2018) focused on early intervention, they ignored dropouts and failures in the dataset. The objective of both the study was to raise the performance of underachievers. However, our study aims at retaining those students who are at risk of not attaining a degree.

2.2.2 Factors of Retention

The performance of classification algorithms is strongly determined by the data used for generating predictive models. Machine learning algorithms use a training dataset for

generating a model that ultimately makes a prediction on an unlabeled test set. The quality of the dataset in terms of the rich set of attributes, as well as the dataset size, impact the performance of the algorithms Jayaprakash et al. (2014). As a result, a large number of studies have incorporated a complex set of features in the dataset, ranging from demographic, social status, economic status, student engagement and interaction, financial standing and more to enrich the dataset and test the theoretical findings on retention.

In this section, we present the focusing the factors used by various studies in predicting dropouts as well as the top predictive factors identified in those studies.

2.2.2.1 Demographic, social and economic factors

Student demographic data is one of the top factors included by various studies in predicting retention. Kovacic (2012) and Martinho et al. (2013) applied data mining techniques using only socio-demographic that were readily available in the institutional database. The dataset included factors such as student age, ethnicity, gender and information about the enrolled program. While Kovacic (2012) used a dataset of only 453 records, Martinho et al. (2013) used a dataset of 1,650 records for training and another dataset of 499 records for testing.

Kovacic (2012) reported an accuracy of 60% and concluded that socio-demographic data alone is insufficient for making predictions, where as the approach by Martinho et al. (2013) yielded a performance of 76%. The difference in performance could be attributed to the difference in the dataset size used in the two studies, thus implicating that not only the quality but also the quantity of the dataset is relevant for machine learning performance.

Delen (2011) reported students age, credit hours and residency status as the top predictive factors of retention. The study by Tamhane et al. (2014) identified math test

scores, ethnicity, and special education needs as top predictor variables. In addition to academic performance, Oztekin (2016), found housing status to be a predictive factor of dropout. Some studies reported factors such as ethnicity (Kovacic, 2012), transferred hours, and proximity to college (Yu et al., 2010) as crucial factors for retention.

These factors are culturally specific indicators of retention. While they may be important in some countries, the factors affecting UAE students are far different. Hence, the result of these studies cannot be applied to our setting.

Theoretical studies on retention have also shown financial standing to be motivation on students persistence in college (Tinto, 1975). However, when using data mining techniques, only a few studies (Delen, 2011; Márquez-Vera et al., 2016, 2013; Perez et al., 2018) have included students' economic status. The principal reason this is the lack of availability of such information since it is often not captured during enrollment. Moreover, none of the studies have reported this as a top predictive factor.

2.2.2.2 Interaction, Engagement and Behaviour data

Bayer et al. (2012) and Aguiar et al. (2014) tested Tinto's social interactionist theory to show that the accuracy of machine-learning algorithms increases with the use of social interaction and engagement data respectively. Both studies reported an increase in performance with the use of engagement data. However, they did not study the statistical significance of the increase.

Bayer et al. (2012) also showed that a reasonable prediction is achieved by the algorithms when student demographic and performance data is used, as opposed to when only interaction data is used. Thus the results of the study revealed that interaction data by itself is not sufficient to predict student success. However, performance data alone can provide a reasonable accuracy of prediction.

Zhang et al. (2010) predicted dropouts using demographic and academic performance data and also applied natural language processing to mine social interaction data from

student discussion boards. The study reported an accuracy of 89.5% using the Naïve Bayes algorithm. However, in our study, we find Naïve Bayes to be one of the most underperforming algorithms among all.

Costa et al. (2017) also augmented their dataset with student engagement data. However, the study focused on pre-processing and optimizing machine-learning algorithms to enhance prediction rates at an early stage. The study showed the importance of pre-processing to improve classification accuracy. However, it did not report whether the use of engagement data enhanced predictions.

EDM techniques have also been applied to study dropouts in online programs (Jayaprakash et al., 2014; Kovacic, 2012; Yukselturk et al., 2014). Kovacic (2012) use only enrollment data and Yukselturk et al. (2014) base their study on data collected from surveys. Online programs have an advantage over traditional on campus studies systems, whereby the systems capture enriched information on student interactions and engagement in the Learning Management Systems (LMS). However, none of these studies leveraged engagement data to predict dropouts. Moreover, Jayaprakash et al. (2014) discarded student interaction data from discussion forums due to the high percentage of missing values.

Behavioural data such as attendance and behavioural incidents are frequently used for predicting high school dropouts. Fernandes et al. (2019) studied the persistence of high school students in Brazil. The study reported attendance as the feature with the highest predictive power for identifying dropouts in the public school.

2.2.2.3 Performance Data

Students' academic performance in college as well as in pre-college is a strong determinant of dropouts in higher education. Performance in high school is directly correlated to performance in university (Danilowicz-Gösele et al., 2017). Moreover, poor academic performance in the first year of studies significantly influences the likelihood of dropping out (Araque et al., 2009; Gershenfeld et al., 2016; Stinebrickner and Stinebrickner, 2014)

In a systematic survey of 30 papers that that use EDM to predict student success, Shahiri et al. (2015) identifed academic data, which includes CGPA and course work assessment scores, as the most significant attributes used to predict student performance in Higher Education. The next most significant attributes included student demographic such as age, gender, as well as pre-enrollment data such as high-school performances and external assessments.

A vast majority of studies have applied EDM techniques using both pre-college and college performance to study the phenomenon of retention. Dekker et al. (2009) and Guarín et al. (2015) showed that using pre-college performance data and first-year college performance data not only enhances the accuracy of the classifier but also that the increase is statistically significant.Raju and Schumacker (2015) used both college performance as well as prior performance to predict droputs and moreover identified first-term GPA, earned hours, student status, and high school GPA as the top factors to predict retention.

Rubiano and Garcia (2015) used data mining methods to show that students' high school performance in physics and biology are strong determinants of success in the first year of college. Oztekin (2016) reported that first term GPA, housing status, and high school score are the strongest determinants of retention. Tamhane et al. (2014) identified math test scores as the top predictive factor of student success.

Although various features with top predictive capabilities have been identified by studies , there is a greater consensus among studies on the predictive power of college performance and pre-college performance. Moreover, other factors such as family history, socio-economic conditions, psychological status and more, although relevant, have a direct or indirect impact on current academic performance in college (Tinto, 1975).

For these reasons we use both high school performance as well as first and second term college performance to train our machine learning models. Moreover, our research shows that high school performance is a strong predictor of graduation when on the pre-college

dataset is used. However, it is the term 1 and term 2 GPA that determines graduation rates in the first year of college.

Socio-demographic and academic performance data is easily obtainable in most institutional database. However, some studies have augmented this data by undertaking surveys (Márquez-Vera et al., 2016, 2013) to collect additional information that is usually not recorded at the time of enrollment or during college life. Márquez-Vera et al. (2013) collect data on socio-economic factors that influence performance such as study habits, friends, number of siblings and more.

Mayra and Mauricio (2018) and Yukselturk et al. (2014) relied solely on survey data for their study. While Yukselturk et al. (2014) studied the readiness and self-efficacy for online programs, Mayra and Mauricio (2018) obtained personal, social, economic and academic data via a survey. Both studies did not include academic performance data stored in the institutional database.

Although a rich dataset that examines retention from various dimensions can be useful, the study by Hoffait and Schyns (2017) concluded that the use of factors such as parents education and family income did not prove to be significant in the prediction of students dropouts in the first year. The study also argued that a student's academic choices and performance are more significant for predicting dropouts than the parents'. Furthermore, obtaining data on the aforementioned factors, via surveys, is a time consuming and costly process. It leads to the reduction in the number instances in the dataset, since surveys are usually optional. For this reason, we base our investigations solely on data available in the institutional database.

Overall, the literature review on the factors involved in student retention show that the most common variables used in predicting student outcome are their current and past performance data as well as demographic data.

2.2.3 Early detection of dropouts

Abu-Oda and El-Halees (2015) and Yu et al. (2010), used data mining algorithms to predict university dropouts using second-year student data. Although it may be assumed that data collected from sophomore students may be enriched with additional college performance, behavioral and interaction data, which in turn would lead to better accuracy, yet, the performance by Yu et al. (2010) yielded an accuracy of 73% and did not surpass other studies that were based on freshmen data. Moreover, our research achieves an accuracy of 79.31% and an AUC score of 88.4% using only pre-enrollment data.

Dropout rates are highest in the first year of studies (Delen, 2011; Hoffait and Schyns, 2017; Tinto, 1975). Early detection of students at risk of dropping out enables HEI to apply strategies to steer students out of the risk. As a result, a large number of studies have focused on early detection of dropouts using enrollment data and first-year performance data. Early identification of potential dropouts is a crucial aspect in our study; therefore we focus on detecting dropouts within the first year of studies.

2.2.4 Dataset size

A major setback in many studies is the use of small dataset to generate prediction models. Márquez-Vera et al. (2016, 2013); Yukselturk et al. (2014); Zollanvari et al. (2017) use a dataset of 419, 670, 189 and 82 respectively to study student success using a data mining approach. While some studies have a small dataset size due to the focus on one discipline of study, others have reduced their dataset size by collecting additional information via surveys.

Aguiar et al. (2014) studied student persistence from one term to another by using 429 records of first-year engineering students' data. The dataset was comprised of demographic, academic performance data and was also enriched with student engagement data. Engagement was considered as the number of logins, submissions and hits on the

eportfolio management system. The study showed that augmenting the dataset with engagement data enhances the performance of the machine-learning algorithms.

Asif et al. (2017) applied classification algorithms on a minimal dataset of 210 records to predict student performance at graduation. Moreover, they classify performance using a multi-class label, which in turn reduces the class distribution of the training set. The study achieved a poor accuracy of 68.85%.

Studies that focus on a single program or course also tend to have smaller datasets. Perez et al. (2018) applied machine-learning algorithms to determine university dropouts in a Computer Science Program. The dataset consists of 802 students enrolled between the academic year 2004 to 2010. It includes demographic data, graduate data, program data such as course grades and CGPA and lastly the financial aid data. The study achieves an AUC-ROC of 94% using the Decision Tree algorithm.

Kovacic (2012) used 453 records of enrollment data to determine the success of students in the Information Systems course in a university. The study achieved a low performance of 60%. Bayer et al. (2012) predicted dropouts using 775 student record of an undergraduate Informatics program. However, they achieve a high accuracy of 92% with the use of ensemble predictors and cost-sensitive learning.

Guarín et al. (2015) use Decision Tree and Naïve Bayes algorithms to predict the loss of academic status in an Engineering program. The study is based on academic and non-academic data of 1,532 students over five years. Naïve Bayes algorithm produced a balanced accuracy of 85% when academic data is included with enrollment data.

Machine learning algorithms are trained using data. A small dataset is counterproductive to a data mining approach as machine learning algorithms can learn better with large datasets. Models based on a small dataset are unreliable and cannot be generalized.

Our research is based on a large dataset of 4056 records, and we achieve high AUC-ROC performance of 92%. We focus on retention in all the programs of study at the college

instead of just one. We rely only on data collected from the institutional database, since augmenting the dataset with additional data that is not available for all students would significantly lower the dataset size. Furthermore, it would exclude potential students who are at-risk of not graduating thereby not serving the purpose of the research.

2.2.5 Balancing Techniques

Imbalanced datasets are common when predicting student success or graduation since there is a large discrepancy between the numbers of students who succeed (also known as the majority or positive class) and those who fail (also known as the minority class). The interest in such datasets in often to accurately predict the negative class. However, an imbalanced dataset leads to misleading results in performance since classification algorithms tend to give preference to the positive class and thereby resulting in a poor prediction of the negative class (Thammasiri et al., 2014).

Surprisingly, only a few of the studies in our literature review report the use of balancing techniques, despite the bias caused in performance.

Researchers overcame the bias of an imbalanced dataset using three main approaches namely, (1) applying cost-sensitive learning, (2) random under-sampling of the majority class or (3) random over-sampling of the minority class. Additionally, the over sampling technique can be a simple duplication of the minority class or creating synthetic observations by using the SMOTE (Synthetic Minority Oversampling Technique) algorithm proposed by Chawla et al. (2002).

Thammasiri et al. (2014) compared all the aforementioned balancing techniques in predicting student success using four different machine-learning algorithms. The study showed that application of any of these techniques improves the performance of the classifiers as compared to using an unbalanced dataset. However, SMOTE algorithm was reported to show the most increase in performance of the algorithms.

In cost-sensitive learning strategy, a cost matrix is used to assign a high cost to misclassification errors of the negative class (Thammasiri et al., 2014). Essentially, the focus of the cost matrix is to minimize the total cost of the misclassifications, thus improving the overall accuracy of the predictions (Ling and Sheng, n.d.). Several researchers (Bayer et al., 2012; Dekker et al., 2009; Guarín et al., 2015; Zhang et al., 2010) used this technique to penalize predictive models and reduce performance bias.

In the random under-sampling technique, observations of the majority class are randomly selected and discarded until the dataset is balanced Thammasiri et al. (2014). It is an effective technique when the original dataset size is large. However, it results in the loss of potentially useful information that could be useful in generating predictive models. Nevertheless, many studies (Aulck et al., 2016; Delen, 2011; Jayaprakash et al., 2014) considered under-sampling as a viable approach to tackle the issue of class imbalance.

Aulck et al. (2016) used under-sampling to balance their dataset of 32,538 records. They evaluated the predictive models using ROC and AUC curves. Delen (2011) applied the under-sampling method to balance their 16,066 records, and further showed that accuracy of the negative class before balancing is almost 50% while balancing improved the performance accuracy to nearly 86.5%.

Random over-sampling is also another technique to balance a dataset. In this method, the minority class observations are duplicated into the dataset until there are enough observations to match the majority class and the dataset is balanced.

A well-known method of oversampling the minority class is the SMOTE algorithm, proposed by Chawla et al. (2002). In this technique, rather than simply duplicating the minority class samples, new artificial samples are generated by using the K-NN algorithm. The SMOTE algorithm has become mainstream and is used by several researchers as a popular data balancing technique Costa et al. (2017); Márquez-Vera et al. (2016, 2013). Márquez-Vera et al. (2013) also showed that the use of SMOTE algorithm improves the

performance of classification models when compared to the baseline and cost-sensitive learning.

Yukselturk et al. (2014) reported a high performance using the k-NN algorithm. However, the dataset used in their study is highly imbalanced, with 20 minority class observations out of the 189. They did not employ any data balancing algorithms or cost-sensitive learning to ensure the results are not biased. Moreover, sensitivity is used as the evaluation measure in the study. The reliability of such a model is questionable since the true positive rate will no doubt be very high since the majority observations are higher in number.

In our research we use the SMOTE balancing technique to resolve performance bias caused by our imbalanced dataset.

Table 2.1 summarizes of all the EDM studies reviewed in this section listing the objective of the study, factors used in prediction, dataset size, balancing technique applied, the machine learning algorithms and evaluation metric used.

Study	Objective	Factors used							Dataset size	Balancing	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic Performance	Engagement/ Interaction	Behaviour	Other				
(Abu-Oda and El-Halees, 2015)	To predict dropout in university	x			x				1290	Oversampling	DT, NB, Association rule	Accuracy
(Aguiar et al., 2014)	To predict performance in an introductory engineering course	x			x	x			429	-	NB, LR, RF	ROC
(Aguiar et al., 2015)	To predict high school dropouts and relevant identify predictors.	x			x			x	11,000	-	LR, RF	ROC, Accuracy, MAE
(Araque et al., 2009)	To detect students at risk in 3 undergraduate programs since 1992	x	x						75,830	-	Statistical methods ,LR and PCA	
(Asif et al., 2017)	to predict the undergraduate performance at the end of four years.			x	x				210	-	DT, IR, K-NN, RF, Clustering	Accuracy, Kappa
(Aulck et al., 2016)	To predict undergraduate dropouts, students who have not completed at least one degree in 6 years of enrollment	x		x	x				69,116	Random Undersampling	LR, K-NN, RF	ROC, RMSE
(Bayer et al., 2012)	To predict student dropout in an undergraduate program	x		x	x		x		775	Cost Matrix	DT, SVM, NB, Ensemble Predictors	Accuracy, TP

Study	Objective	Factors used							Dataset size	Balancing	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic Performance	Engagement/ Interaction	Behaviour	Other				
(Costa et al., 2017)	To predict likely failures in programming course	x	x		x	x			262 - distance learning; 161 – on campus	SMOTE	NB, SVM, DT, NN	F-Measure
(Danilowicz-Gösele et al., 2017)	To study graduation predictors of university students				x				12, 315	-	Regression	
(Dekker et al., 2009)	To predict students who are likely their diploma in a 3 year program.	x		x	x				648	Cost matrix	DT, RF, NB	Accuracy
(Delen, 2011)	To identify students who are likely to dropout after their first year study in a public university	x	x	x	x				16,066	Random Undersampling	Ensemble predictors, DT, NN, LR, SVM, RF	Accuracy
(Djulovic and Li, 2013)	To predict freshman retention in university students	x	x	x	x				7800	-	DT, NB, NN, IR	Accuracy, Recall and Precision
(Fernandes et al., 2019)	To predict academic performance of high school students in Brazil	x			x				485872	-	GBM	ROC
(Gershenfeld et al., 2016)	To study the impact of first term GPA on retention and graduation rates	x		x	x				1,947			
(Guarín et al., 2015)	To predict undergraduate dropout in first year of college	x	x		x				1532	Cost matrix	NB, DT	Accuracy

Study	Objective	Factors used							Dataset size	Balancing	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic Performance	Engagement/ Interaction	Behaviour	Other				
(Hoffait and Schyns, 2017)	To predict first year students who are most likely to drop out	x	x	x	x				11,496	-	RF, LR, NN	Accuracy
(Huang and Fang, 2013)	To predict the performance in an engineering course				x				323	-	LinR , MLP, SVM	Accuracy
(Jayaprakash et al., 2014)	To predict students at risk of failing an online course course	x		x	x	x			9,938 training size; 5,212 testing dataset size	Random Oversampling of minority and Undersampling of majority class	LR, SVM, DT, NB	Recall, FP rate and precision
(Kovacic, 2012)	To predict students success in an Information Systems course	x	x						453	-	DT, LR	Accuracy
(Lakkaraju et al., 2015)	To predict dropout in high school	x			x		x		200000	-	Ensemble predictor, RF, LR, DT, SVM	ROC, Precision, Recall, New metric
(Márquez-Vera et al., 2013)	To predict school dropouts in first year of high school	x	x		x		x	x	670	SMOTE, Cost Sensitive Learning	IR, DT	Accuracy

Study	Objective	Factors used							Dataset size	Balancing	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic Performance	Engagement/ Interaction	Behaviour	Other				
(Márquez-Vera et al., 2016)	To predict course dropouts in high school	x	x		x		x	x	419	SMOTE	NB, SVM, K-NN, DT, Genetic Algorithm	GM, Accuracy
(Martinho et al., 2013)	To predict risk groups of student dropout	x	x						1650 training; 499 testing	-	NN	Accuracy
(Miguéis et al., 2018)	To predict final academic performance	x	x		x				2,459	-	Ensemble Predictors, RF, DT, SVM, NB	Accuracy
(Natek and Zwilling, 2014)	To predict the final grade (low, medium, high) in an undergraduate course	x	x		x				106	-	DT	Accuracy
(Perez et al., 2018)	To predict dropouts and identify important factors leading graduation in a computer science program	x	x	x	x				802	Dataset is balanced	DT, LR, NB	ROC
(Raju and Schumacker, 2015)	To predict graduation and identify relevant factors for predicting student graduation	x		x	x				22099	-	DT, LR, NN	ROC
(Rubiano and Garcia, 2015)	To predict at risk students in first-year Systems Engineering program	x	x		x				932	-		

Study	Objective	Factors used							Dataset size	Balancing	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic Performance	Engagement/ Interaction	Behaviour	Other				
(Yu et al., 2010)	To predict retention using student data in the second year of college	x		x	x				6,690	-	LR, DT, SVM	Accuracy
(Yukselturk et al., 2014)	To predict dropout in an online program	x						x	189	-	NB, DT, K-NN, NN	Sensitivity
(Zhang et al., 2010)	To predict dropout due to poor performance in university using data mining and NLP	x			x	x			4223	Cost Matrix	NB, SVM, DT	Accuracy
(Zollanvari et al., 2017)	To predict a undergraduate student's GPA							x	82	-	Joint Probablity	Accuracy, specificity and sensitivity

Table 2.1: Summary of all studies

Chapter 3

Methodology

In this research we follow the CRISP-DM (cross-industry process for data mining) methodology Shearer (2000), which provides a structured and organized framework for performing data mining tasks. The CRISP-DM approach consists of six phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

The following sections discuss the tasks performed in each phase. Fig. 3.1 shows an overview of the tasks performed in each phase of the CRISP-DM methodology

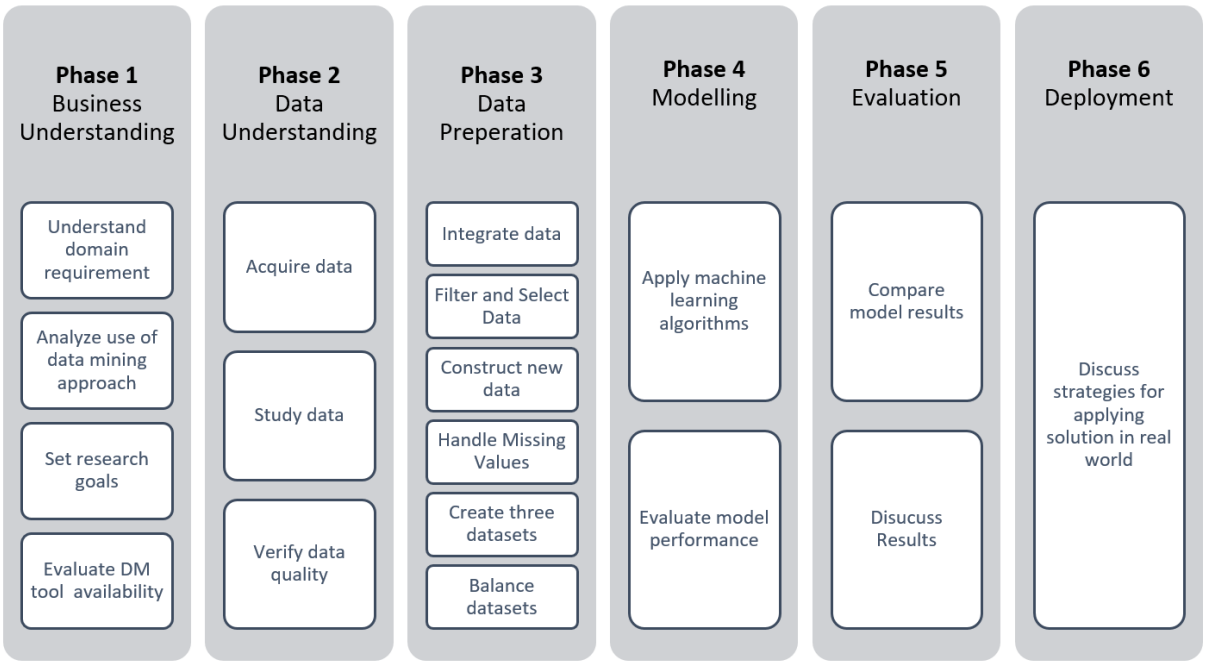


Figure 3.1: CRISP-DM Phases

3.1 Phase 1 - Business Understanding

This research addresses the issue of student retention in a UAE based HEI using a data mining approach. The institution has two campuses in Dubai and offers six bachelor degree programs of study, namely Business, Applied Media, Computer and Information Science, Engineering, Health Science and Education. Each program offers several concentrations, which the students choose in year 2 or 3 of their studies based on the program requirements.

Nearly 6000 students enroll each term into the various programs, out of which almost 1300 are new enrollments. The education of all students is funded either federally or through a local sponsor. All the students in the college belong to the same ethnic and cultural background and form a homogeneous group.

Admission to the college undergraduate program is subject to the students' performance in high school, IELTS exam and a Common Entry Placement Assessment(CEPA) in Math and English. A student must achieve a CEPA Math score of 170 and above and a CEPA English score of 180 and above to secure a place in the bachelor program.

Failing to meet the entry requirements, places students in the foundation program for a year where they improve their English and Math skills after which they may enroll in the bachelor programs.

As per institutional requirements, a student must graduate within a nominal six-year time frame from a bachelor degree program. Failing to do so would cause the student to prematurely end their education without earning a degree. Throughout this research, we use the term dropout to refer to students who have left college before graduating in the program they have enrolled in.

Student retention is the ability to institution to successfully retain students until graduation. It is a critical issue whereby one of the key performance indicators of the college is to maximize retention. Nearly 30% of students who enrolled in the year 2011 did not graduate on time. This research aims at providing a solution to the institution through early detection of students likely to drop out by applying data mining techniques. Our research will enable the institution to take preemptive measures to reduce dropout rates.

A drop out may occur in one of three ways - 1) student officially withdraws from college, 2) student is academically dismissed due to poor performance 3) student does not re-enroll in the following terms without formally withdrawing. However, a student who stops studies for a term and re-enrolls in the following term to complete his/her studies is not considered as a dropout. Furthermore, a student who changes his/her program of study is considered a dropout in the first program he/she enrolled in and new enrollment in the other program of study.

The dependent variable in our research is a binary attribute labeling the student as graduated or not. Our research aims to provide answers to three research questions that were outlined in section 1.2.1. Here we describe the plan on how these research questions will be answered.

Research Question 1: Can machine-learning algorithms effectively predict retention/dropouts among a homogeneous population of students?

To answer this research question, we generate predictive models to classify dropouts. We use five standard classification algorithms (Decision Tree, Naïve Bayes, Logistic Regression, Deep Learning and SVM) and five ensemble algorithms (Gradient Boosted Trees, Random Forest, AdaBoost, Bagging and Voting).

Due to the stochastic nature of machine-learning algorithms, predictive models cannot perform with 100% accuracy. Therefore, an effective model is one with a reasonable performance in a given domain. Based on reviewed literature, we evaluate the effectiveness of our predictions against a threshold requirement of 75% AUC.

Research Question 2: How early can we predict potential dropouts using machine learning?

To answer this research question, we use three datasets. The first dataset includes pre-college data such as demographic information and pre-college performance data. The second dataset contains all the data from the first dataset along with term1 performance, while the third dataset includes all data of the previous datasets and is additionally augmented with term2 performance. The performance of the algorithms with each dataset is compared using a pair-wise t-test to determine if there is a significant increase in performance or merely due to chance.

Research Question 3: Which attributes are the top predictors of retention?

To answer this research question, we first analyze the attributes by feature weights to study its relevance with regards to the class label. We begin by calculating the feature weight of each attribute using information gain, gain ratio, gini index, correlation and chi-squared statistic. We then identify the top predictors of retention by examining the nodes fo the Decision Tree model. We also compare the weights assigned to each attribute by various algorithms such as SVM, Logistic Regression, Random Forest, Gradient Boosted Trees and report the topmost predictive factors for each dataset.

3.2 Phase 2 - Data Understanding

In the Data Understanding phase, we acquire and explore the data to study relationships between attributes and identify quality issues such as outliers and missing values. We develop a strategy in this phase to deal with the identified problems. Data selection strategies and pre-processing requirements are further studied at this stage and later implemented in Data Preparation phase. The following section describes the datasets used in our research.

3.2.1 Data Collection

One of the main challenges of EDM is the acquisition and integration of data, which is dispersed across multiple systems and in various formats (Romero and Ventura, 2013).

In this study we use two main datasets – the enrollment and graduation. The enrollment data is acquired from the institutional database by extracting all the enrollment reports between the academic year 2011 to 2014, inclusive (excluding the summer term, since new enrollments do not take place in summer). This dataset contains 84 features that includes student demographic data, high school performance scores, IELTS performance scores, CEPA scores the current term, term GPA, program of enrollment and other enrollment details. The enrollment dataset consists of 22,000 records in total that are extracted in eight separate Excel files, one for each term.

The graduation reports extracted from the system contains information of all graduates since the inception of the college. This dataset contains 22,680 records and includes demographic data, graduation year, program or graduation, specialization and degree attained.

In this phase we study both datasets carefully to identify features with missing values and outliers which are handled in the next phase.

3.3 Phase 3 - Data Preparation

The Data Preparation phase is the most critical and labor intensive of all phases. This phase is further subdivided into two steps, the data integration and initial preprocessing, which is performed using Microsoft Access, and the additional preprocessing, performed using Rapid Miner. The two sub stages of the Data Preperation phase are described in the following sections.

3.3.1 Data Integration and Initial Pre-processing

In the initial pre-processing phase, we begin by merging the enrollment records and then filtering and selecting the relevant records. We then integrate the enrollment and graduation datasets and construct new attributes. Figure 3.2 shows the data integration and initial pre-processing tasks performed using Microsoft Excel and Microsoft Access.

The next sections describe the data selection, integration and new data construction in detail.

3.3.1.1 Data Selection

The enrollment records are stored in eight different excel files, one for each enrollment term for the academic years 2011 to 2014. We begin our initial pre-processing by merging all the eight files into one. We then select the relevant records as explained in the next sections.

Filter Undergraduate Enrollments - Our study is based on retention in undergraduate degree programs. However, the dataset includes enrollment records of all programs such as Foundations, Diploma, Higher Diploma and Bachelor degree. We extract only the Bachelor degree enrollment records. Furthermore, the Foundation program does not lead to graduation and the Diploma and Higher Diploma programs are currently phased out.

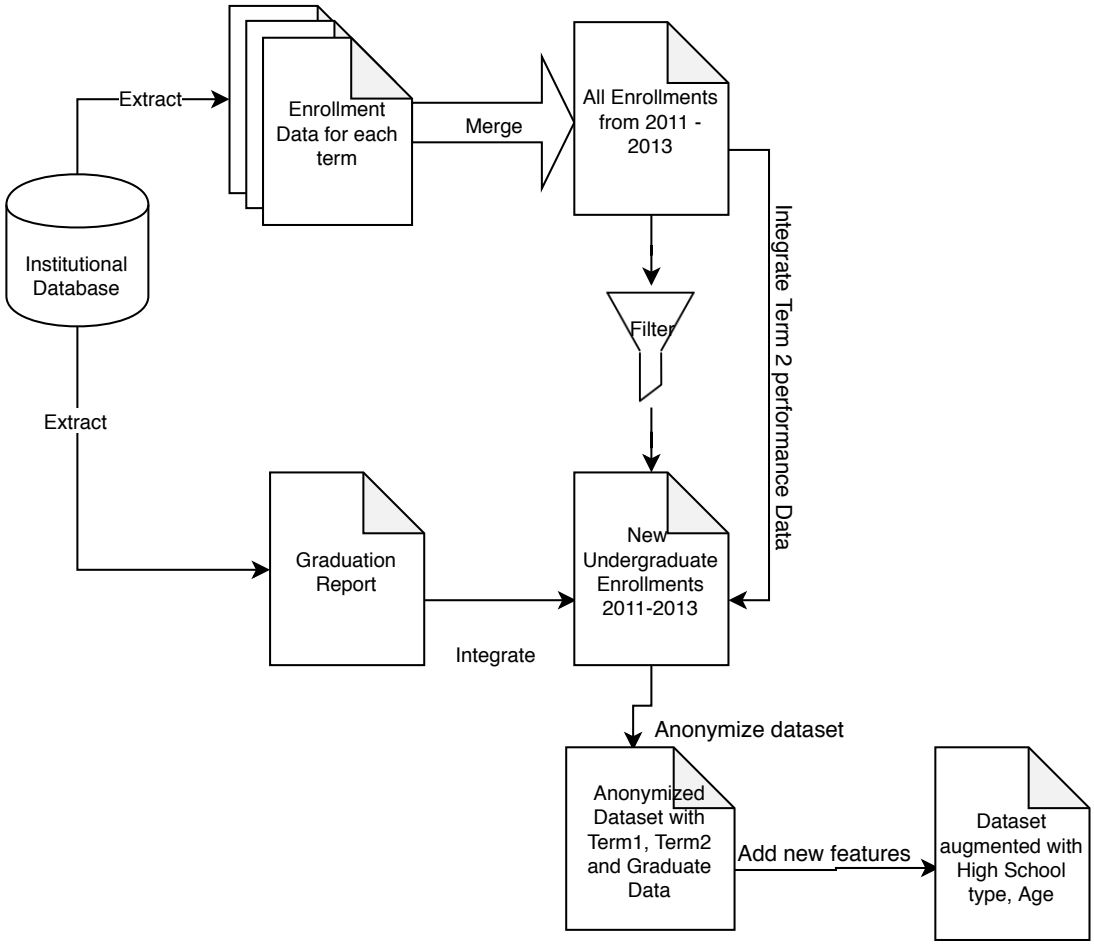


Figure 3.2: Initial Preprocessing in MS Excel and MS Access

Filter New Enrollments - The enrollment dataset includes new and returning enrollment records in each term. However, our study requires only the new enrollments in each term to predict graduation or dropouts. Therefore, we filter and extract the new enrollments by matching the current term with enrollment term.

Moreover, we also filter out new enrollments who have withdrawn, failed due to attendance, academically dismissed or have a term GPA of 0. After preliminary analysis of the dataset, we discover that the enrollment dataset contains duplicate student entry. These cases are of students who change their program of study. We consider these cases as dropout in one program and new enrollment in another program.

Remove non-graduates enrolled in 2014 – We filter out those students who enrolled in academic year 2014 but did not graduate yet since these students have not com-

pleted their nominal 6 year period of on-time graduation and may potentially graduate on time. However, we include the new enrollment records of 2014 that graduated.

3.3.1.2 Data Integration

Integrate Term 2 GPA - The enrollment dataset contains only the Term 1 GPA of the enrolled students. To get the term 2 GPA, we import our dataset to MS Access and perform join operations between the cleaned up the dataset and the old enrollment records to extract the term 2 GPA and integrate it in our dataset. Some students do not have term 2 GPA since they would have dropped out after term 1 or withdrawn for that term.

Integrate Graduation Data – The graduation report is stored as a separate dataset. We apply inner joins in MS Access to match the enrollment record with the graduation record to determine if newly enrolled students have graduated in the program they enrolled in. We create a new feature, the class label, called Graduated, which takes binary values of 'Y' or 'N'.

3.3.1.3 Construct new data

We impute two new features called High School type and age.

High School type is imputed from the high school name and is used as an alternative to represent the socioeconomic background of the students. Private schools in Dubai are very costly with a fee ranging from 20,000 AED to 100,000 AED per year. Students that attend private high schools typically belong to families with a higher socioeconomic standard. Our data consists of an attribute called high school name. Accordingly, we determine whether the student has attended a private or public high school by cross-listing the high school name with a list of all private high schools in Dubai[1].

[1]http://www.dubaifaqs.com/schools-dubai.php

Age at enrollment can have an impact on drop out rates (Dekker et al., 2009). We compute student age at enrollment by using the students' date of birth and the enrollment date.

3.3.1.4 Anonymize and clean up the Dataset

Our dataset has 84 attributes in total. The dataset is anonymized by removing personal information about the student, which includes student name, contact information, passport details. Furthermore, all irrelevant attributes, which add no value to our study or were completely missing, are also removed. The dataset is left with 36 attributes.

3.3.2 Additional Pre-processing

The quality of the dataset determines the performance of the data mining algorithms (Delen, 2011). Therefore, pre-processing is a crucial step that is needed to ensure the success of the algorithms as well as to ensure the validity of the results. We import the initially pre-processed data into Rapid Miner and further pre-process it to prepare for the application stage. Four main pre-processing tasks are performed in Rapid Miner, which include handling missing values, feature selection, creating sub-datasets, balancing the dataset. Figure 3.3 shows the tasks performed with Rapid Miner to pre-process the data, generate and apply the predictive models.

3.3.2.1 Missing Values

To handle features with missing values we use two techniques. First, any feature that has more than 20% of missing values, such as CEPA Writing Task2 and Completed Volunteering hours, is excluded from the dataset. Second, the missing values of all other attributes (less than 20% of missing values) are imputed using k-NN algorithm with a value of k set to 3. No action is taken for Term 2 GPA since a missing value indicates

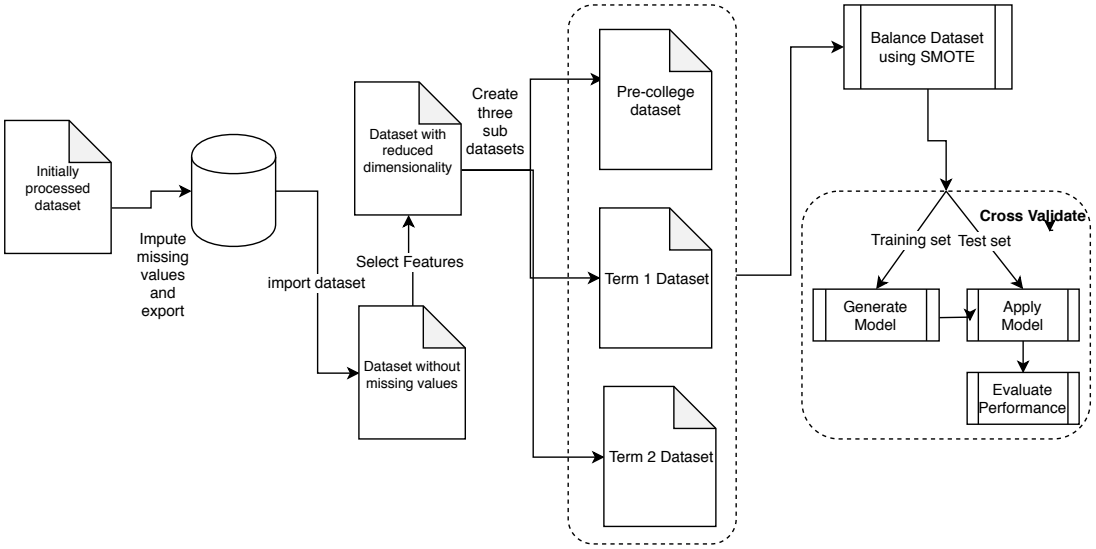


Figure 3.3: Rapid Miner Process

that the student did not register for term 2 and dropped out in term 1. Table 3.1 shows a summary of all attributes with missing values.

Feature	**% missing**	**Action Taken**
Term 2 GPA	4.01%	No Action
CEPAW Task 2	93.08%	Exclude Attribute
Completed Volunteering Hours	48.14%	Exclude Attribute
HS Average	4.35%	Impute Missing Value
High School Type	3.93%	Impute Missing Value
HS Stream	0.44%	Impute Missing Value
HS English	17.89%	Impute Missing Value
HS Math	19.03%	Impute Missing Value
HS Arabic	18.89%	Impute Missing Value
CEPA	4.56%	Impute Missing Value
CEPAW	5.54%	Impute Missing Value
CEPA Math	19.25%	Impute Missing Value
IELTS Writing	13.40%	Impute Missing Value
IELTS Reading	13.40%	Impute Missing Value
IETLS Listening	13.40%	Impute Missing Value
IELTS Speaking	13.50%	Impute Missing Value
IETLS Band	12.41%	Impute Missing Value
IELTS Overall	12.65%	Impute Missing Value

Table 3.1: Attributes with missing values

Imputing the missing values using k-NN algorithm is a time-consuming process that

would require to be performed for each experimentation run. Therefore, to increase the efficiency of the experiments, we export the dataset with the imputed values and used later import it in another Rapid Miner process for the next pre-processing and application tasks.

3.3.2.2 Feature Selection

High dimensionality of a dataset leads to many problems such as overfitting of the models in which case the model cannot be generalized (Aguiar et al., 2014). We further reduce the features of the dataset by selecting only those that are related to student demographic data, enrollment data, pre-college performance and college performance. Moreover, any feature that is used to generate new attributes, such as catalog terms, High School name, birth date are excluded from the study.

The resultant dataset consists of the features shown in Table 3.2

We use five statistical methods to study the relevance of each independent variable with regards to the class label. The methods include information gain, gain ratio, gini index, correlation and chi-squared statistic. Each methods assigns a feature weight to the independent variable. A high feature weight indicates a higher relevance to the class label. We rank each feature from 1-19 according to the weight assigned by the algorithms. The higher the ranking, the higher the relevance of the feature. Table 3.3 shows the list of all the feature rankings sorted by the most highly ranked feature.

Some of the least relevant features are found to be CEPA English and High School type, while the most relevant features appear to be the college and high school performance scores.

3.3.2.3 Create three sub-datasets

The pre-processed dataset is divided into three sub datasets, which are used to determine the earliest stage potential dropouts can be predicted. The three datasets are described

Type	Sub Type	Attribute	Data Type	Description	Range of values
Demographic Data	Personal	Age	Numeric	Age at enrollment ranging from 17 to 45	16 - 45 (Avg 20.85, Deviation - 4.05)
		Gender	Binary	Gender of student - M for Male and F for Female	M = Male (57%) F = Female (43%)
	Education	Coll	Nominal	The division / college student is enrolled in- Includes Business, CIS,	BU = Business (45%) ET = Engineering (26%) IT = Info. Technology (14%) CT = Communication (7%) HS = Health Science (4%) ED = Education (4 %)
		HS Type	Binary	Type of High School - Public or Private	Public (72%) Private (28%)
		HS Stream	Nominal	Chosen stream in High School. Values include General, Arts, Science etc..	Arts (36%) General (29%) Diploma (25%) Other (10%)
		CatalogTerm	Numeric	The term that the student is enrolled in	For example: 201110 denotes first term of academic year 2011
Pre-college performance	High School	HS Avg	Numeric	High School Average percentage	44 - 99.4 (Avg - 78.42, Deviation - 9.611)
		HS Math	Numeric	Math Score ranging from	40 - 100 (Avg -77, Deviation - 12.08)
		HS English	Numeric	English score	40 - 100 (Avg -77, Deviation - 9.8)
		HS Arabic	Numeric	Arabic score	42 - 100 (Avg - 77, Deviation - 12)
	Standardized Tests	IELTS Band	Numeric	IELTS Overall Band	1.5 - 8 (Avg - 5.3, Deviation - 0.5)
		IELTS L	Numeric	IELTS Listening Band	3.5 - 9 (Avg - 5.5, Deviation - 0.736)
		IELTS W	Numeric	IELTS Writing Band	1 - 8 (Avg - 5, Deviation - 0.55)
		IELTS S	Numeric	IELTS Speaking Band	4 - 8.5 (Avg 5.7 , Deviation - 0.685)
		IELTS R	Numeric	IELTS Reading Band	2 - 8.5 (Avg - 5, Deviation - 0.627)
		CEPA	Numeric	CEPA English Score	123 - 210 (Avg - 175, Deviation - 11)
		CEPA Math	Numeric	CEPA Math Score	90 - 210 (Avg - 150, Deviation - 16)
College	Performance data	Term 1 GPA	Numeric	Semester GPA achieved in Term 1	0 - 4 (Avg - 2.75, Deviation - 0.944)
		Term 2 GPA	Numeric	Semester GPA achieved in Term 2	0 - 4 (Avg - 2.77, Deviation - 1.024)
Class Label		Graduated	Binary	The class label used for prediction which indicates whether the student graduated or not	Y = Yes (74%) N = No (26%)

Table 3.2: Dataset Features

Feature	Chi Squared Statistic	Correlation	Gain Ratio	Information Gain	Gini Index	Average
Term2_GPA	19	19	19	18	18	18.6
Term1_GPA	18	18	18	19	19	18.4
HS_Avg	17	17	10	17	17	15.6
HS_Arabic	14	14	14	13	13	13.6
HS_Math	16	16	6	15	15	13.6
CEPA MATH	12	13	12	12	12	12.2
Gender	13	15	5	14	14	12.2
HS_Stream	15	10	4	16	16	12.2
IELTS R	10	9	16	10	10	11
HS_English	11	12	9	11	11	10.8
IELTS W	9	11	17	8	8	10.6
IELTS Band	8	8	8	9	9	8.4
IELTS L	7	7	7	5	5	6.2
CEPW	6	6	3	7	7	5.8
Age	4	4	11	3	3	5
IELTS S	2	3	13	2	2	4.4
Coll	5	2	1	6	6	4
CEPA	1	1	15	1	1	3.8
HS_Type	3	5	2	4	4	3.6

Table 3.3: Feature Weights

below:

1. **Pre-college dataset:** consists of demographic data, pre-college performance (High School, IELTS, CEPA scores)

2. **College Dataset 1:** Pre-college dataset + Term 1 GPA score + program of study

3. **College Dataset 2:** College Dataset 1 + Term 2 GPA score (exclude records with missing values of Term 2 GPA)

3.3.2.4 Balance Datasets

The dataset at this stage is imbalanced with the majority class (graduated students) representing 74% of the observations while the minority class (dropouts) representing

only 26%. A model generated by the machine-learning algorithms using an imbalanced dataset would be produce misleading results. Therefore, balancing the dataset reduces the bias caused by the majority class and improves the performance of machine algorithms (Guarín et al., 2015; Thammasiri et al., 2014)

In educational data mining, balancing techniques have proven to be useful when predicting dropouts in Mexican high schools (Márquez-Vera et al., 2016). While there are many ways of balancing a dataset, specifically, in this research we have chosen the SMOTE algorithm proposed by (Chawla et al., 2002).

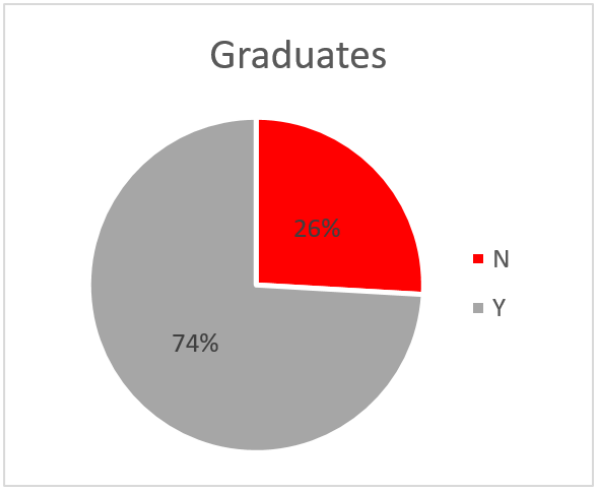
The SMOTE algorithm is popular technique used for balancing a dataset used in several studies (Costa et al., 2017; Márquez-Vera et al., 2016). Using this technique, we first extract the minority class observations from the dataset. Then a random minority class observation is selected along with a random neighbour, using the k-nearest neighbour algorithm, with the value of k set to 5. A new observation is synthetically generated between the two selected observations. This process is repeated until our dataset is balanced.

3.3.3 Descriptive statistics

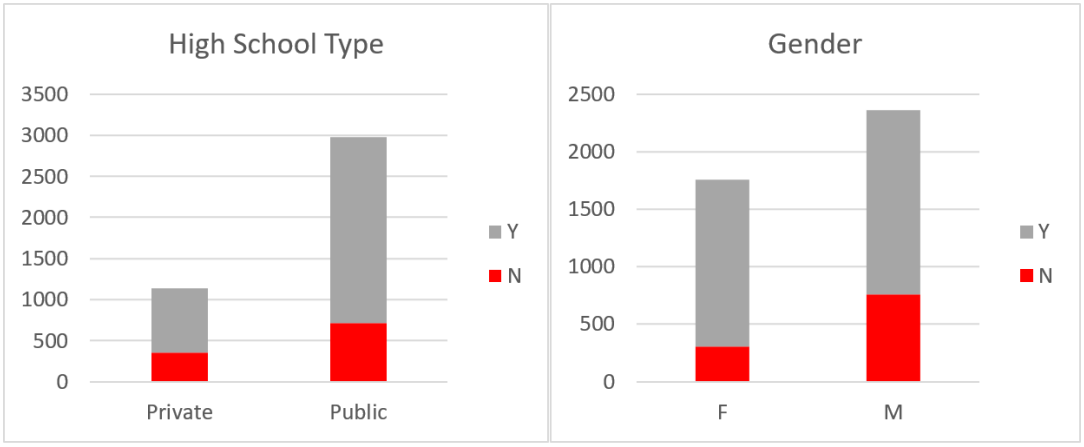
The resultant dataset after the data preparation phase has 4,056 records and includes enrollment data of two campuses from the period of 2011 to 2014. Figure 3.4 shows a descriptive statistics of the number of graduates by gender, High School type, year and undergraduate programs.

3.4 Phase 4 - Modelling

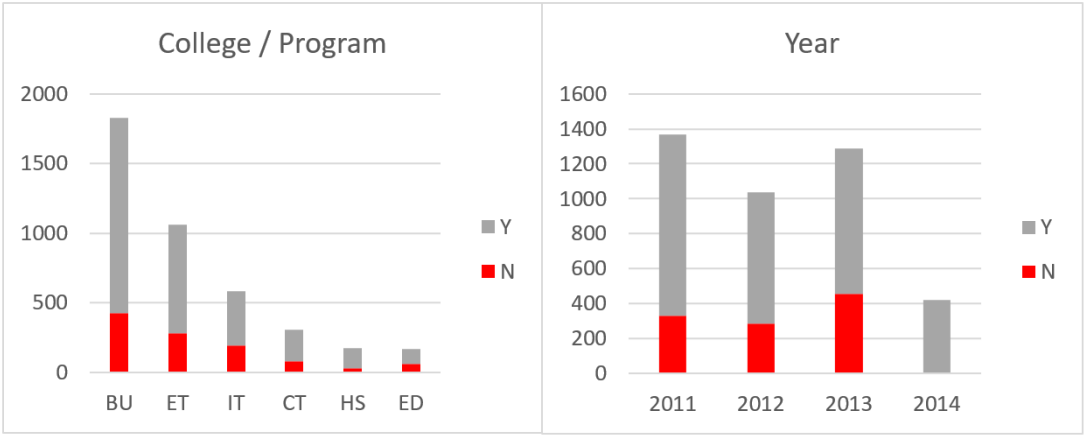
In the Modelling phase, we select the supervised machine learning algorithms that will be used to generate the predictive models. This research uses six standard classification



(a) Graduates



(b) Graduates by highschool type and gender



(c) Graduates by program and year

Figure 3.4: Descriptive Statistics of Dataset

algorithms, namely, Decision Tree, Naïve Bayes, k-NN, Logistic Regression, SVM and Deep Learning. In addition, the study also employs the following five ensemble algorithms to achieve a robust prediction – Random Forest, Voting, Bagging, AdaBoost and Gradient Boosted Trees.

The parameters of some algorithms are adjusted to optimize its performance. The models are evaluated using several metrics: Area under the curve (AUC) of the Reciever Operator Characteristic (ROC) curve, F-measure, True Positive Rate (TPR), True Negative Rate (TNR) and Accuracy. Top performing models are also compared to determine if the difference in performance is statistically significant.

3.4.0.1 Training and Validation

The classification models are produced using 10-fold cross-validation. A stratified sampling technique is used to split the original dataset into ten subsets while preserving the ratio of the minority and majority samples. The machine learning algorithm is trained using nine subsets, while testing is performed using the remaining subset. This process is repeated ten times by holding out another subset and training with the remaining nine. The final performance is reported as the average of all the iterations.

3.4.1 Standard Machine Learning Algorithms

3.4.1.1 Decision Tree

Decision Tree classifier generates a predictive model as a hierarchical structure that is easy to understand and interpret (Witten et al., 2011). The tree is built by recursively splitting the dataset into nodes using a specific feature, which produces the least impurity. A pure node represents all the class labels of one class, which in our context means an outcome of graduate or not. The process is repeated until a node is either pure or too

small to be split. The nodes are split using one of many criterions such as information gain, gain ration or gini index.

A decision tree is highly prone to overfitting. Therefore it is important to prune the tree. Pruning reduces the unnecessary splits in the tree thereby reducing its complexity. It is achieved by setting several parameters such as maximum depth of a tree, minimum number of instances per node and the minimum number of instances required to split at the node, minimal gain, which is the threshold of impurity.

Decision Tree Parameters

A decision tree produces optimal results when its parameters are fine-tuned. Hence, we use a parameter optimization operator in Rapid Miner to adjust the spit criterion, minimal gain and confidence parameters of the tree to maximize its performance. Results of the optimization are shown in Table 3.4.

Dataset	**Split criterion**	**Minimal gain**	**Confidence**	**AUC**
Pre-college dataset	Gini-index	0.019	0.22	74.7%
College dataset 1	Gini-index	0.064	0.5	81.4%
College dataset 2	Gini-index	0.046	0.34	82.1%

Table 3.4: Optimization parameters of Decision Tree

Discretization of Values

Discretization is the process of transforming a continuous variable into discrete and ordered grouped intervals. Since our dataset contains several continuous variables such as age, GPA, high school, IELTS and CEPA performance scores we discretize these variables into bins, specifically for the Decision Tree algorithm. Age is discretized using specific bins with five years interval, starting from 16-20 till 41-45. All other numeric attributes are discretized in 5 bins. Table 3.5, 3.6 and 3.7 shows the intervals of discretized values.

	Age	GPA	HS Average	HS English and HS Math	HS Arabic
Range 1	16-20	-∞ - 0.928	-∞ - 55.080	-∞ - 52	-∞ - 59.82
Range 2	21-25	0.928 - 1.696	55.08 - 66.16	52 - 64	59.82 - 69.64
Range 3	26-30	1.696 - 2.464	66.16 - 77.24	64 - 76	69.64 - 79.46
Range 4	31-35	2.464 - 3.232	77.24 - 88.32	76 - 88	79.46 - 89.28
Range 5	36-40	3.232 - ∞	88.32 - ∞	88 - ∞	89.28 - ∞
Range 6	41-45				

Table 3.5: Discretized ranges of Age, GPA and High School scores

	Term 1 GPA	CEPA English	CEPA Math
Range 1	-∞ - 0.928	-∞ - 140.40	-∞ - 114
Range 2	0.928 - 1.696	140.40 - 157.80	114 - 138
Range 3	1.696 - 2.464	157.80 - 175.20	138 - 162
Range 4	2.464 - 3.232	175.20 - 192.60	162 - 186
Range 5	3.232 - ∞	192.60 - ∞	186 - ∞

Table 3.6: Discretized ranges of Term1 GPA and CEPA scores

	IELTS Listening	IELTS Reading	IELTS Writing	IELTS Speaking	IELTS Band
Range1	-∞ - 4.600	-∞ - 3.30	-∞ - 2.400	-∞ - 1.700	-∞ - 2.800
Range 2	4.600 - 5.700	3.300 - 4.600	2.400 - 3.800	1.700 - 3.400	2.800 - 4.100
Range 3	5.700 - 6.800	4.600 - 5.900	3.800 - 5.200	3.400 - 5.100	4.100 - 5.400
Range 4	6.800 - 7.900	5.900 - 7.200	5.200 - 6.600	5.100 - 6.800	5.400 - 6.700
Range 5	7.900 - ∞	7.200 - ∞	6.600 - ∞	6.800 - ∞	6.700 - ∞

Table 3.7: Discretized ranges of IELTS score

3.4.1.2 k-Nearest Neighbour (k-NN)

k-NN is a distance-based classification algorithm. It uses a distance metric to identify k nearest neighbours of a given observation and uses a voting system to classify the observation into its respective class (Witten et al., 2011). The distance metric used is the Euclidean Distance which is computed by measuring the distance between two vectors p and q as shown in eq.(3.1)

$$d(p,q) = \sqrt{\sum_{n=1}^{n}(q_i - p_i)^2} \tag{3.1}$$

Using parameter optimization process in Rapid Miner, we determine the best value of k is 4.

3.4.1.3 Naïve Bayes

Naïve Bayes is a probabilistic classifier and uses the most basic classification algorithm that predicts the outcome of classification using the Bayes Theorem (Witten et al., 2011). This algorithm assumes that each attribute in the training set is independent of each other. The Naïve Bayes operator in Rapid Miner uses the Gaussian probability theorem.

3.4.1.4 Logistic Regression

Logistic regression is a popular algorithm used for binary classifications. In this algorithm, a sigmoid function coefficient is estimated from the training dataset (Witten et al., 2011). The Logistic Regression model links the predictor variables to probabilities through (3.2)

$$P(x) = \frac{(e^{b_0+b_1 x})}{1 + e^{b_0+b_1 x}} \tag{3.2}$$

Where $P(x)$ is probability estimate whose output is between 0 and 1 and e is the base of natural log. The coefficients of the logistic regression (b_0 and b_1) are estimated from the training dataset using the maximum likelihood estimation, which is a common learning algorithm. The best coefficients are the are the ones that would predict a value very close to 1 for the positive class (graduated) and a value very close to 0 (dropout) for the negative class.

We use default parameters provided by Rapid Miner for the Logistic Regression algorithm. Numeric attributes are standardized to have zero mean and unit variance and collinear columns are removed to avoid overfitting of the model.

3.4.1.5 Deep Learning

Deep Learning is based on a multi-layer feed-forward artificial neural network that is trained using back-propagation. The network can contain a large number of hidden layers

consisting of neurons with rectifier activation function, which is implemented using the library in H_2O.

3.4.1.6 Support Vector Machines (SVM)

SVM consist of an input, a layer of trained support vectors, and a classification output. SVMs use a training dataset to find a minimum, optimal distance between cases from two different classes or subsets of the dataset (Provost and Fawcett, 2013).

The SVM and Deep Learning algorithm cannot deal with nominal attributes. Hence, we convert all the nominal attributes to numerical values for these classification algorithms.

3.4.2 Ensemble Predictors

A recent trend in classification techniques is to combine multiple machine-learning algorithms to produce better and more robust predictions (Seni and Elder, 2010). These are call ensemble predictors. The most popular ensemble predictor amongst studies in our reviewed literature is the Random Forest algorithm. However, some studies (Aguiar et al., 2014; Delen, 2011; Lakkaraju et al., 2015; Miguéis et al., 2018) have leveraged the prediction power of multiple ensemble predictors in addition to individual machine learning algorithms to generate robust predictions.

There are two types of ensemble algorithms – Bagging and Boosting. The bagging technique uses combines the predictions of several machine learning algorithms using an averaging technique. Examples of bagging techniques include random forest and voting. In the boosting technique, the prediction of an algorithm is enhanced incrementally to improve the classification of the previous classifier by learning from its mistakes. Examples of boosting techniques include AdaBoost and Gradient Boosted Trees.

In this research, we use both bagging and boosting ensemble techniques to combine the top performing algorithms and generate more robust results. We use five main ensemble predictors – Random Forest, Voting, Bagging, AdaBoost and Gradient Boosted Trees.

3.4.2.1 Random Forest (RF)

Random forest algorithm aggregates the predictions of multiple decision trees (Kuhn and Johnson, 2013). The final classification is based on the tallying the all the votes of all the trees in the forest.

3.4.2.2 Voting

The voting algorithm uses multiple standard algorithms to build classification models and assigns the average of all predicted values to make a prediction (Kuhn and Johnson, 2013). We use the Naïve bayes, Decision Tree and Logistic Regression as the base learners in the voting ensemble.

3.4.2.3 Bagging

The bagging technique gives each classification an equal weight and determines the combined prediction (Kuhn and Johnson, 2013). We use bagging with the decision tree algorithm using 10 iterations.

3.4.2.4 AdaBoost

Boosting is the process of improving the performance of weak classifiers. AdaBoost algorithm builds a model using the training data, then improves that model in subsequent attempts by tweaking previous models misclassifications. This process is repeated until the maximum number of iterations are reached. AdaBoost works by assigning a high

weight to instances that are difficult to classify. In the next iteration, the classifiers focus on the difficult instances (Kuhn and Johnson, 2013).

We use the AdaBoost algorithm to boost the performance of the decision tree algorithm using 10 iterations.

3.4.2.5 Gradient Boosted Trees (GBT)

The GBT model uses a forward-learning approach to predict the outcome of classification using gradually enhanced estimates (Kuhn and Johnson, 2013). Multiple decision trees are constructed to produce a collection of weak prediction models. The errors in each model are analyzed and focused on by subsequent models. At the end, all the predictions are combined by assigning weights to each prediction. Our model combines the prediction of 20 trees.

3.4.3 Classification accuracy measures

The aim of our research is to effectively predict dropouts using a dichotomous class label that classifies the graduates (positive class) and dropouts (negative class). Our primary interest is to accurately predict the dropouts while not misclassifying the graduates. Inaccurate classification results in poor utilization of resources in dropout interventions for students who are likely to graduate, whilst also missing out on students who are actually at risk of dropping out.

We use four main evaluation metrics to measure the performance of our predictive models– Accuracy, Recall, TNR (True Negative Rate), TPR (True Positive Rate), AUC (Area Under the Curve) of ROC (Receiver Operator Characteristic) curve and F-Measure.

The classification algorithms classify each instance of the testing data in into two classes, Y (graduated) or N (did not graduate). The four possible classifications of the instances are captured in a confusion matrix shown in Table 3.8.

Graduated	**Actual (Y)**	**Actual (N)**
Predicted (Y)	True Positive	False Positive
Predicted (N)	False Negative	True Negative

Table 3.8: Confusion Matrix

The matrix summarizes the actual and predicted values:

True Positive (TP) are the number of positive class observations that are classified correctly.

False Negative (FN) are the number of positive class observations classified incorrectly.

True Negative (TN) are the number of negative class observations classified correctly.

False Positive (FP) are the number of negative class observations classified incorrectly.

Each of the performance evaluation metric is briefly explained below:

Accuracy is popular metric used by many researchers (Abu-Oda and El-Halees, 2015; Bayer et al., 2012; Dekker et al., 2009; Delen, 2011; Guarín et al., 2015; Hoffait and Schyns, 2017; Huang and Fang, 2013; Kovacic, 2012). It computes the percentage of correct classifications in the matrix by computing the total correct classifications over all the classifications as shown in eq. (3.3).

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{3.3}$$

Accuracy is an unreliable metric when used with an imbalanced dataset as it can result in a performance bias. Since a balancing technique is applied to our dataset, we use accuracy to test the overall performance the models.

Specificity, also known as the **True Negative Rate (TNR)**, measures the number of negative class observations that are correctly classified. Within the context of this research, specificity would be the percentage of dropouts that are correctly classified as dropouts as shown in eq. (3.4)

$$TNR = \frac{TN}{(TN + FP)} \tag{3.4}$$

A **False Positive Rate (FPR)** denotes the percentage of dropouts that are misclassified as is computed as shown in eq. (3.5).

$$FPR = 1\text{–}TNR \tag{3.5}$$

Precision, also known as **Positive Predictive Value (PPV)** measures the number of positive class observations among all positive predictions as shown in eq. (3.6).

$$PPV = \frac{TP}{(TP + FP)} \tag{3.6}$$

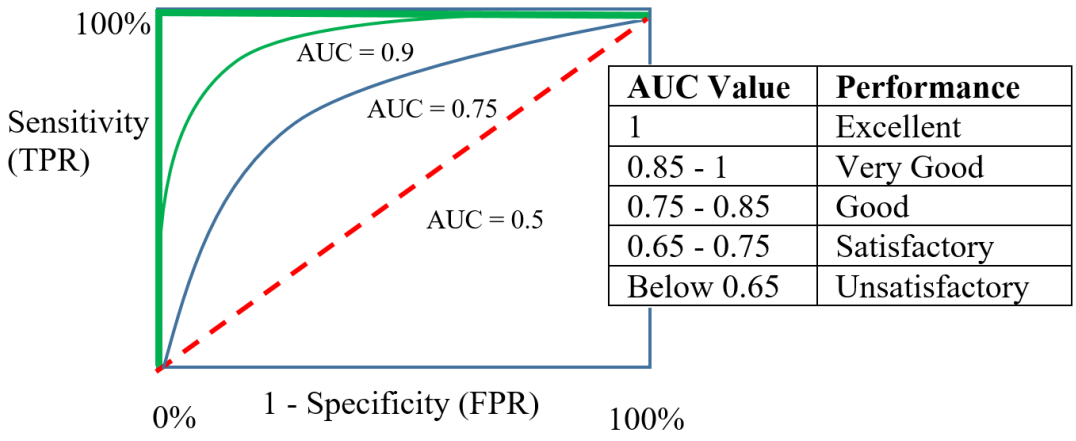
Sensitivity, also known as **Recall** or **True Positive Rate (TPR)**, is a measure of the number of positive class observations that are correctly classified as shown in eq. (3.7). Within the context of this study, specificity denotes the percentage of graduates that are classified correctly.

$$TPR = \frac{TP}{(TP + FN)} \tag{3.7}$$

ROC (Receiver Operator Characteristic) Curve displays the performance of the classification model by plotting the sensitivity and specificity against a threshold (Fawcett, 2006). ROC analysis provides a holistic evaluation measure instead of just relying on precision, sensitivity or specificity (Bowers et al., 2012).

ROC curves focus on correctly identifying the graduates (TP) while lowering the misclassification of the dropouts (TN). It is plotted with the Sensitivity (TPR) in the y-axis and 1- specificity (FPR) in the x-axis. The area under the curve (AUC) is a metric between 0 - 1 that is used to evaluate the performance of the model. An AUC value of 1 indicates that the performance of the algorithm is excellent with no false positives and no false negatives. An AUC value of 0.5 indicates that the model has classified an equal number of true positives and false positives. Therefore, within the context of this research, a higher value of AUC indicates that the model has correctly classified a large number of graduates and a large number of dropouts.

Figure 3.5 shows an example of ROC curves with Very Good (AUC = 0.9), Good (AUC = 0.75) and Unsatisfactory performance curves. The range of values have been chosen with respect to the reviewed literature.



AUC Value	Performance
1	Excellent
0.85 - 1	Very Good
0.75 - 0.85	Good
0.65 - 0.75	Satisfactory
Below 0.65	Unsatisfactory

Figure 3.5: ROC Curves examples

3.5 Phase 5 - Evaluation

In the Evaluation phase, the results obtained from each model using each of the three datasets are studied and compared. Chapter 4 discusses the results.

Chapter 4

Results and Discussion

The purpose of this research is to use a data mining approach to predict undergraduate students at risk of dropping out without earning a degree. In addition, we seek to determine the earliest stage when an effective prediction is possible and to identify the top predictive factors of retention at that stage.

Three research questions were presented in Chapter 1. In this section, we discuss our experimental findings and answer each question.

4.1 Research Question 1

Can machine-learning algorithms effectively predict retention/dropouts among a homogeneous population of students?

To answer this research question, we generate predictive models on three datasets using five standard classification algorithms as well as five ensemble algorithms. The performance of each model is evaluated to determine

if the algorithm performs effectively, with an AUC of at least 75%. We begin our discussion by explaining the effect of discretization on the decision tree and the behaviour of the k-NN algorithm that led us to exclude these algorithms from our experiments.

4.1.1 Decision Tree Classifier

In the preprocessing stage, we discretized the numerical attributes such as age, high school grades, CEPA and IELTS score as well as the term 1 and term 2 GPA. Details of the discretization are available in Section 3.4.1.1 The performance of the decision tree with and without discretization is shown below in Figure 4.1.

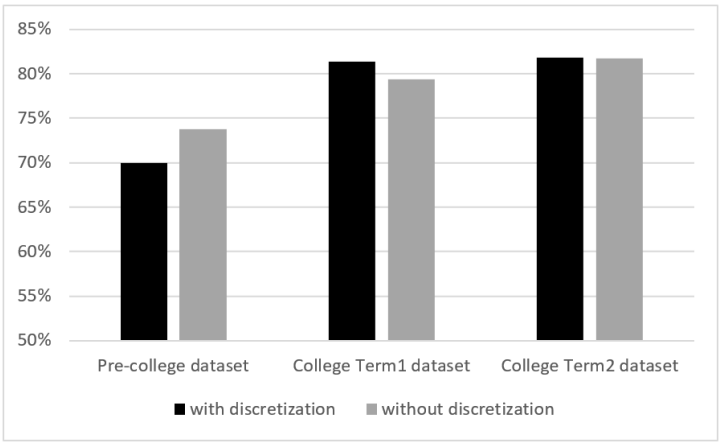


Figure 4.1: Decision Tree Performance

Table 4.1 shows the outcome of the pairwise t-test conducted to study the statistical significance of the difference in performances of the Decision Tree with and without discretization.

For the pre-college dataset, the performance of the Decision Tree is

reduced by 3.8% with the use of discretization. A pairwise t-test shows that the decrease in performance is indeed significant with α=0.001. However, the difference in performance for the College Term 1 and Term 2 datasets are not significant, with an α value of 0.645 and 0.893 respectively.

	Pre-college dataset	**College Term1 dataset**	**College Term2 dataset**
AUC with discretization	70%	81.40%	81.80%
AUC without discretization	73.80%	79.40%	81.70%
Significance of difference	α = **0.001**	α = **0.645**	α = **0.893**

Table 4.1: Significance of difference in Decision Tree performances

The result shows that discretization lowers the performance of the Decision Tree for the pre-college dataset and has no significant effect on performance for the Term 1 and Term 2 datasets. Therefore, for the rest of the experiments, we use the decision tree without discretization.

4.1.2 k-NN Classifier

The k-NN algorithm behaves differently than all other classifiers in predicting retention, with a consistent performance across all three datasets. Figure 4.2 shows the performance of the classifier on all the three datasets.

The k-NN algorithm classifies dropouts better, with a true negative rate (TNR) of 78-81%, than graduates with a TPR of 61% - 65%. This result may be attributed to the fact that the minority class observations of the dataset (dropouts) are synthetically generated by applying the SMOTE

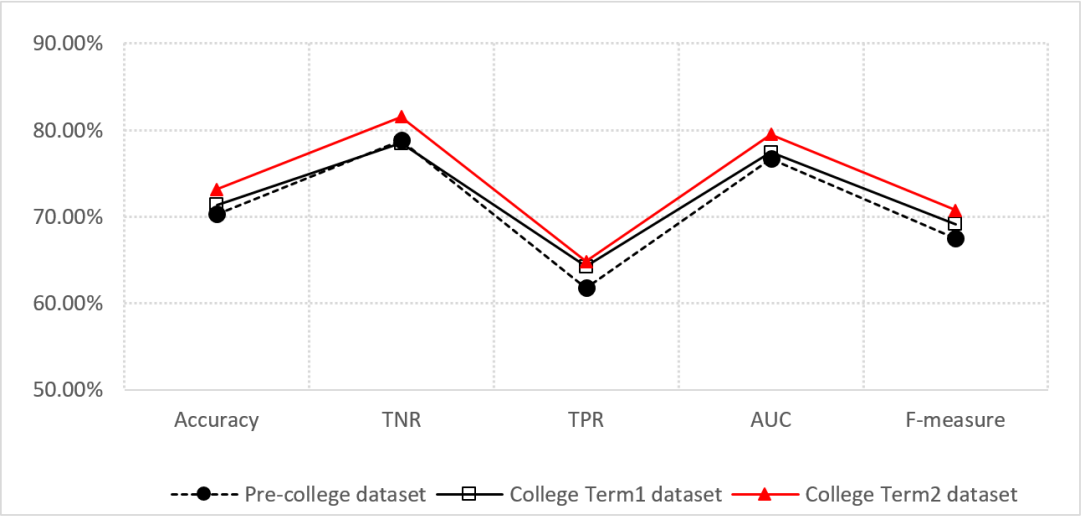


Figure 4.2: k-NN classifier performance on all datasets

algorithm, using k-NN. As a result, the classification of the negative class is better than the classification of the positive class. Thus, we consider the k-NN results to be unreliable for the negative class.

Due to the unreliable behaviour of the k-NN this model is excluded from the classifier performance comparisons.

4.1.3 Standard Classifier Results

In this section, we compare the results of the five standard classifiers for each dataset to determine if they are effective in predicting dropouts.

4.1.3.1 Pre-college dataset

Table 4.2 shows the performance of the standard classifiers on the pre-college dataset, with the top performance for each metric highlighted.

	Accuracy	TNR	TPR	AUC	F-Measure
Decision Tree	**71.59%**	67.17%	76.02%	73.80%	**72.80%**
Naïve Bayes	67.22%	**77.10%**	57.34%	73.90%	63.60%
Logistic Regression	67.92%	70.22%	65.63%	**74.40%**	67.16%
SVM	67.09%	72.64%	61.53%	73.30%	65.13%
Deep Learning	62.80%	46.40%	**79.19%**	70.16%	68.02%

Table 4.2: Standard algorithm performance on pre-college dataset

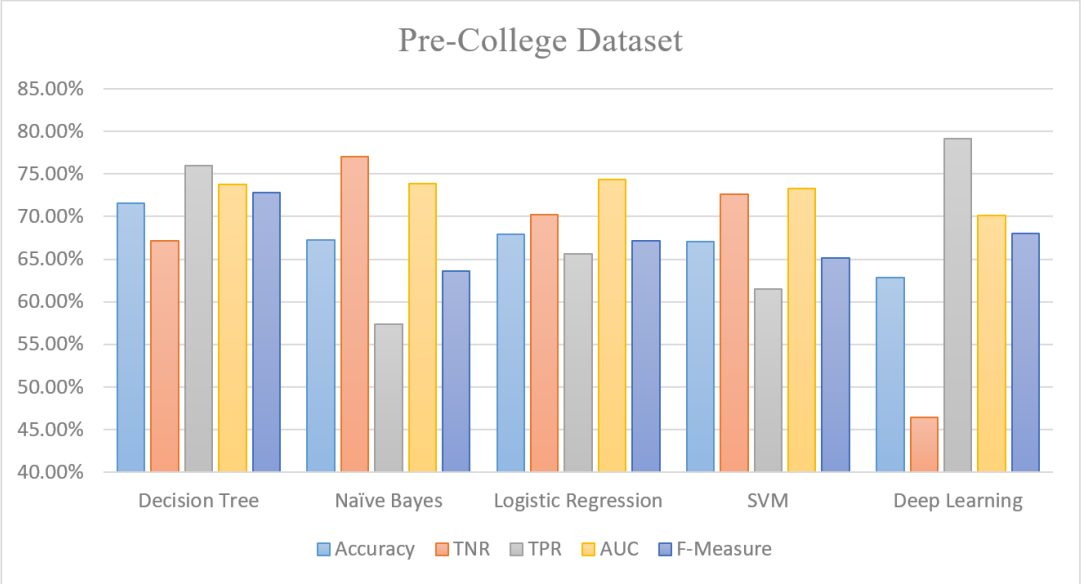


Figure 4.3: Standard algorithm performance on pre-college dataset

As shown in Figure 4.3, the AUC of the ROC curve of all the algorithms used for the pre-college dataset ranges from 70% to 74.4%. Since our threshold for selecting an effective algorithm is 75% AUC, we find that none of the standard classifiers perform successfully for the pre-college dataset.

However, the decision tree has the best accuracy and F1-score of 71.5% and 72.8% respectively. The DT model and the logistic regression models balance the predictions of the positive and negative class with similar values of TPR and TNR. The deep learning model is good at predicting the graduates with TPR at 79.1%, nevertheless, it has very low predictive

power in classifying dropouts with a TNR of 46.4%. The SVM and Naïve Bayes on the other hand are able to predict the dropouts with a rate of 72% and 77% respectively; however, the prediction of graduates is on the lower end of 61% and 57% respectively.

4.1.3.2 College Term 1 Dataset

Table 4.3 shows the performance of the standard classifiers on the College Term 1 dataset, with the top performance in each metric highlighted.

	Accuracy	**TNR**	**TPR**	**AUC**	**F-Measure**
Decision Tree	75.77%	70.54%	**81.00%**	79.40%	76.96%
Naïve Bayes	73.97%	**77.49%**	70.45%	81.80%	73.01%
Logistic Regression	77.20%	75.82%	78.57%	**84.80%**	77.51%
SVM	**77.31%**	75.46%	79.23%	84.70%	**77.77%**
Deep Learning	69.68%	61.14%	78.21%	77.50%	72.07%

Table 4.3: Standard algorithm performance on college term 1 dataset

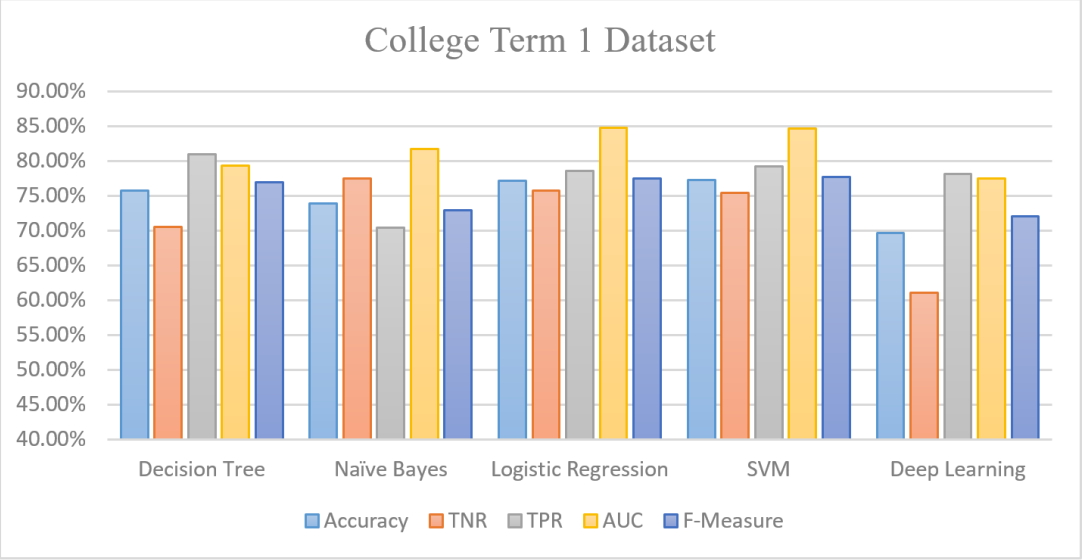


Figure 4.4: Standard algorithm performance on college term 1 dataset

As shown in Figure 4.4, all the standard algorithms perform effectively with the college term 1 dataset with an AUC score above the threshold

requirement ranging from 77% to 84.8%. The logistic regression and SVM algorithms perform the best, achieving accuracies of up to 77.3%. Both algorithms predict the graduate and dropouts equally well with TPR of 79.2% and TNR of 75.4%. The Decision Tree and Naïve Bayes algorithms attain the next best performance. The deep learning algorithm is still the weakest in predicting dropouts and has a big discrepancy between the TPR and TNR rates.

4.1.3.3 College Term 2 Dataset

Table 4.4 shows the performance of the standard classifiers on the College Term 2 dataset, with the top performance in each metric highlighted.

	Accuracy	**TNR**	**TPR**	**AUC**	**F-Measure**
Decision Tree	**80.41%**	**88.40%**	72.41%	81.70%	**78.70%**
Naïve Bayes	75.23%	76.93%	73.53%	83.33%	74.81%
Logistic Regression	77.42%	76.51%	78.34%	**86.00%**	77.62%
SVM	77.95%	76.38%	79.52%	85.80%	78.27%
Deep Learning	75.20%	67.01%	**83.39%**	82.90%	77.08%

Table 4.4: Standard algorithm performance on college term 2 dataset

All the standard algorithms have improved their effectiveness with the college term 2 dataset and achieved an AUC score above 80%. Again, the logistic regression and SVM algorithms have the highest AUC score of 86%, and overall accuracy is 77.9%.

Although the accuracy of the decision tree algorithm is the highest amongst all other algorithms at 80.4%, the AUC score is lower than the other algorithms. This indicates that the decision tree was not able to

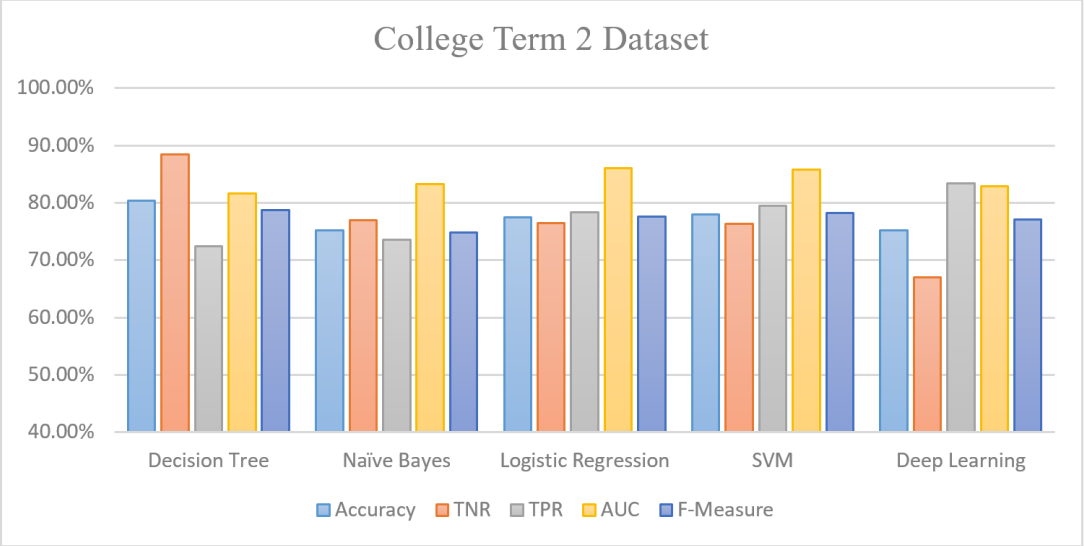


Figure 4.5: Standard algorithm performance on college term 2 dataset

achieve a high true positive and false negative score. The decision tree is the best at predicting the dropouts with the highest TNR of 88.4%, but the prediction of graduates is at the lower end of 72.4%. The deep learning algorithm is still the weakest at predicting dropouts with a rate of 67%.

The overall accuracy of the Naïve Bayes classifier has improved when the college term 1 and term 2 datasets are used. However, there is no improvement in prediction of the dropouts which remains constant at 77% across all three datasets. The addition of college data improves the ability of the Naïve Bayes algorithm to predict graduates from 57% to 73%.

4.1.4 Ensemble Classifier Results

In this section, we compare the results of the five ensemble classifiers for each dataset to determine if they are effective in predicting dropouts.

4.1.4.1 Pre-college dataset

Table 4.5 shows the performance of the ensemble classifiers on the pre-college dataset, with the top performance in each metric highlighted.

	Accuracy	TNR	TPR	AUC	F-Measure
Random Forest	70.35%	73.75%	66.94%	77.10%	69.21%
Gradient Boosted Trees	**79.31%**	72.35%	**86.27%**	**88.40%**	**80.66%**
Voting	70.54%	**75.00%**	66.09%	71.90%	69.15%
AdaBoost	71.23%	66.74%	75.72%	71.20%	72.47%
Bagging	73.02%	69.10%	76.87%	79.50%	74.02%

Table 4.5: Ensemble algorithm performance on pre-college dataset

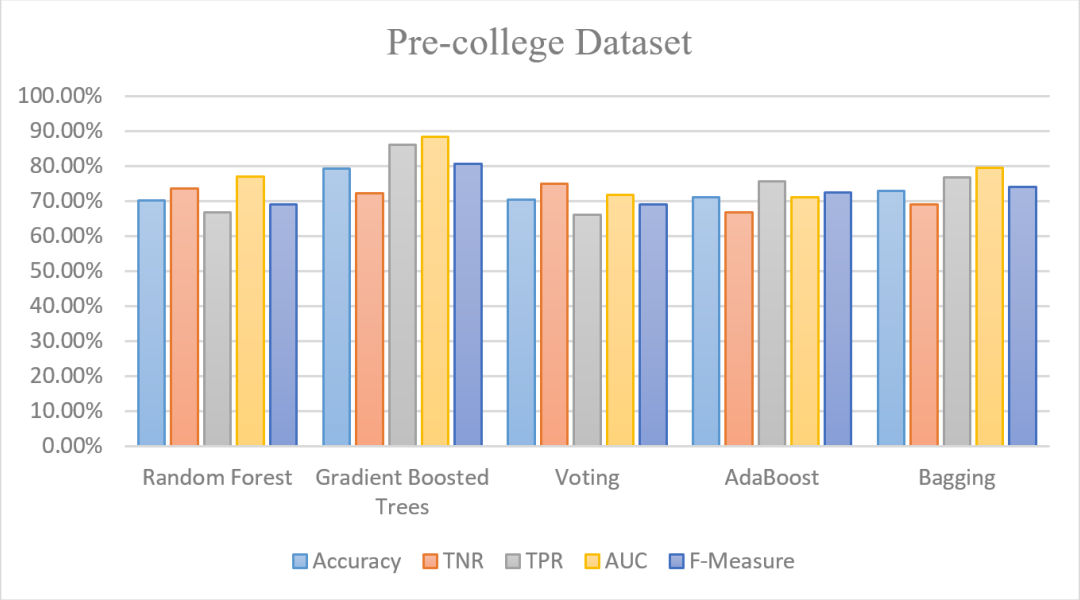


Figure 4.6: Ensemble algorithm performance on pre-college dataset

As shown in Figure 4.6, all the ensemble algorithms have performed better than the standard algorithms when the pre-college dataset is used. All algorithms except the Voting and AdaBoost classifiers meet the threshold requirement of 75% AUC. The Gradient Boosted Tree classifier has the best accuracy and AUC score of 79.3% and 88.4% respectively. However, it is better at predicting graduates, at 86.2%, than at predicting dropouts at

72.3%. The Voting algorithm and Random Forest are the best at predicting dropouts with a rate of 75% and 73.7% respectively.

4.1.4.2 College Term 1 Dataset

Table 4.6 shows the performance of the ensemble classifiers on the college term 1 dataset, with the top performance in each metric highlighted.

	Accuracy	TNR	TPR	AUC	F-Measure
Random Forest	77.79%	73.00%	82.57%	85.00%	78.84%
Gradient Boosted Trees	**82.10%**	**79.59%**	**86.04%**	**90.10%**	**83.35%**
Voting	77.26%	74.93%	79.59%	77.70%	77.77%
AdaBoost	75.62%	70.48%	80.77%	78.60%	76.80%
Bagging	76.69%	71.63%	81.75%	83.40%	77.79%

Table 4.6: Ensemble algorithm performance on college term 1 dataset

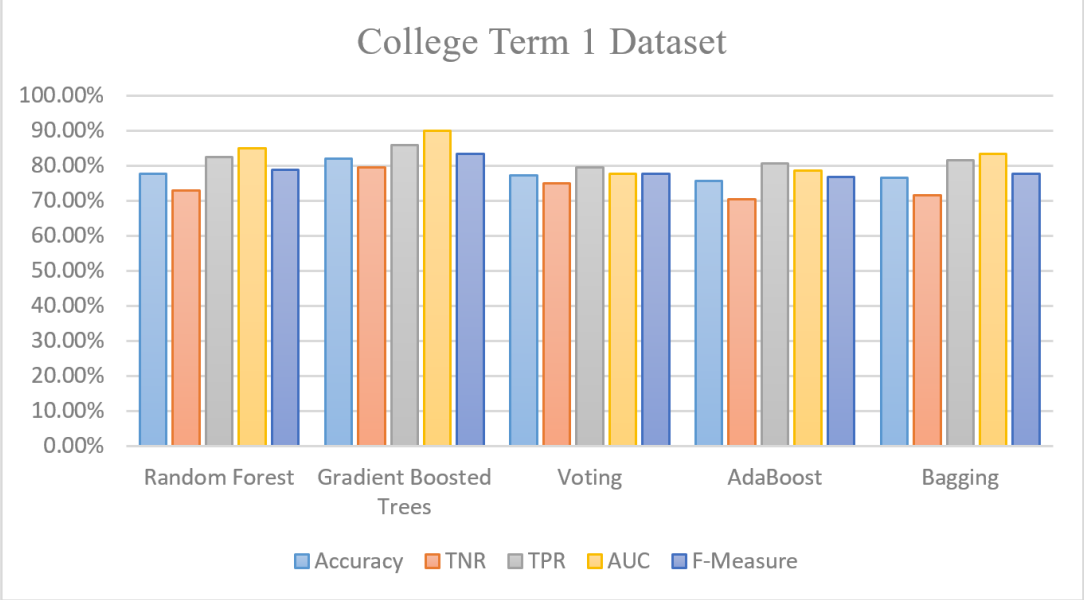


Figure 4.7: Ensemble algorithm performance on college term 1 dataset

As shown in Figure 4.7, the performance of the Gradient Boosted Tree is still the best for the College Term 1 dataset. It has an AUC score of 90.10% with an accuracy of 82.10%. TPR and TNR are also balanced

at 86% and 80% respectively. All other ensemble classifiers also perform better when College Term 1 Dataset is used. The poorest performance is that of the AdaBoost classifier with an accuracy of 75.6%. It also shows the least predictive capability in identifying dropouts at a rate of 70.4%.

4.1.4.3 College Term 2 Dataset

Table 4.7 shows the performance of the ensemble classifiers on the college term 2 dataset, with the top performance in each metric highlighted.

	Accuracy	**TNR**	**TPR**	**AUC**	**F-Measure**
Random Forest	81.16%	82.67%	79.65%	88.80%	80.88%
Gradient Boosted Trees	**84.75%**	83.32%	**86.17%**	**92.20%**	**84.97%**
Voting	78.96%	80.83%	77.10%	81.60%	78.56%
AdaBoost	80.21%	88.40%	72.02%	83.90%	78.44%
Bagging	82.01%	**88.47%**	75.56%	89.10%	80.76%

Table 4.7: Ensemble algorithm performance on college term 1 dataset

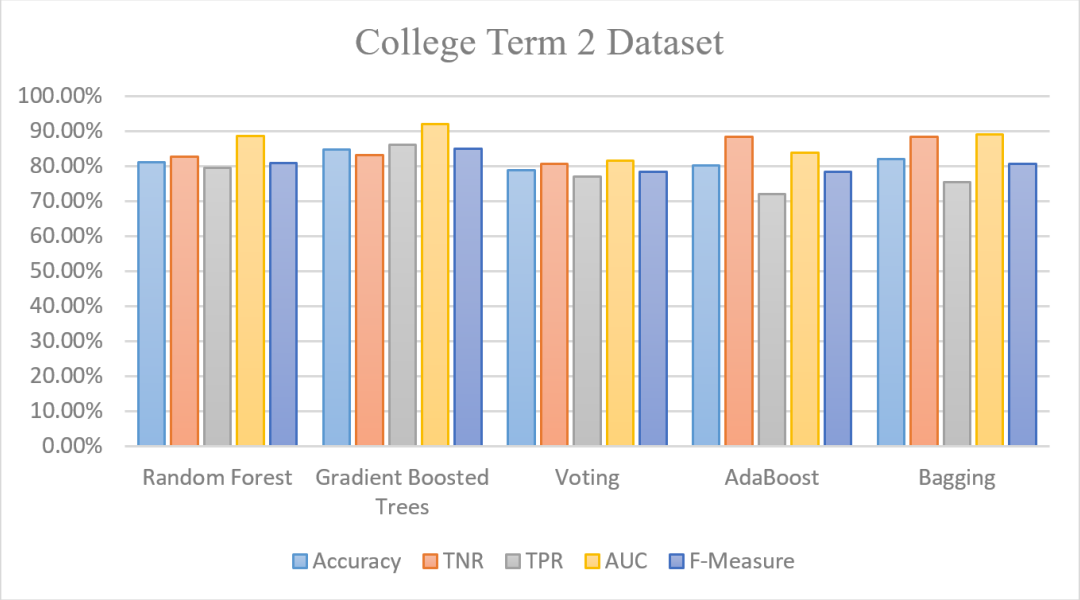


Figure 4.8: Ensemble algorithm performance on college term 2 dataset

The Gradient Boosted Trees classifier is once again the most effective when the College Term 2 Dataset is used. As shown in Figure 4.8, it achieves an AUC score of 92.2% with an accuracy of 84.7%. All other classifiers have also improved in their performance with the lowest accuracy of 78.9% achieved by the Voting algorithm. AdaBoost and Bagging algorithms are the best at predicting dropouts at 88.4%, while the lowest predictive capability of dropouts is that of the Voting Algorithm at 80.8%.

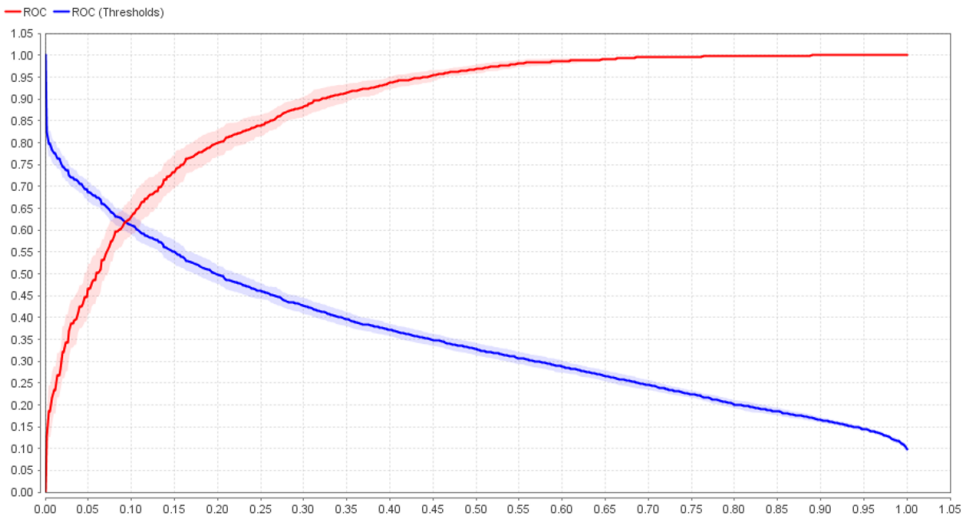
Overall, our results show that all standard algorithms, as well as few of the ensemble algorithms, are weak at making predictions with the pre-college dataset and do not meet our threshold requirement, AUC score of 75%. Nevertheless, dropouts can be predicted effectively with the precollege dataset using Gradient Boosted Trees, Random Forest and Bagging ensemble algorithms.

The ensemble algorithms have performed better than all standard algorithms across all the three datasets. The Gradient Boosted Trees classifier consistently outperforms all other ensemble and standard classifiers, making it the most effective.

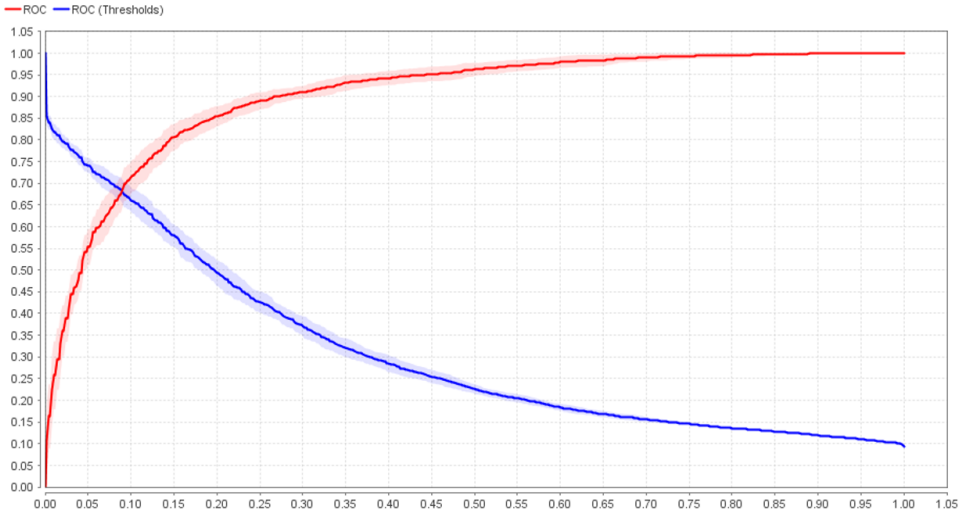
The AUC for the ROC curve for the Gradient Boosted Trees for all the three datasets is shown in the Figure 4.9.

4.2 Research Question 2

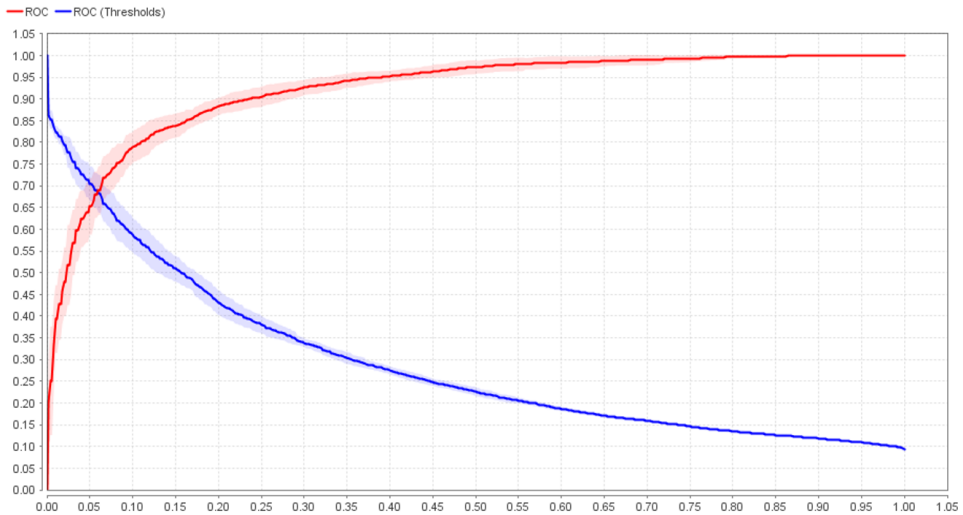
How early can we predict potential dropouts using machine learning?



(a) Pre-college dataset



(b) College Term 1 dataset



(c) College Term 2 dataset

Figure 4.9: ROC curves of the Gradient Boosted trees for all datasets

Beyond determining which classifier can accurately predict students at risk of dropping out, the objective of our study is also to make this prediction at an early stage. We answer our research question by examining the performance of all classifiers across all the datasets. We also compare the difference in performance across the datasets using a pair-wise t-test to determine if the increase in performance is significant or merely due to chance.

4.2.1 Performance by dataset – standard algorithms

Figure 4.10 shows the AUC score of all standard algorithms across the three datasets.

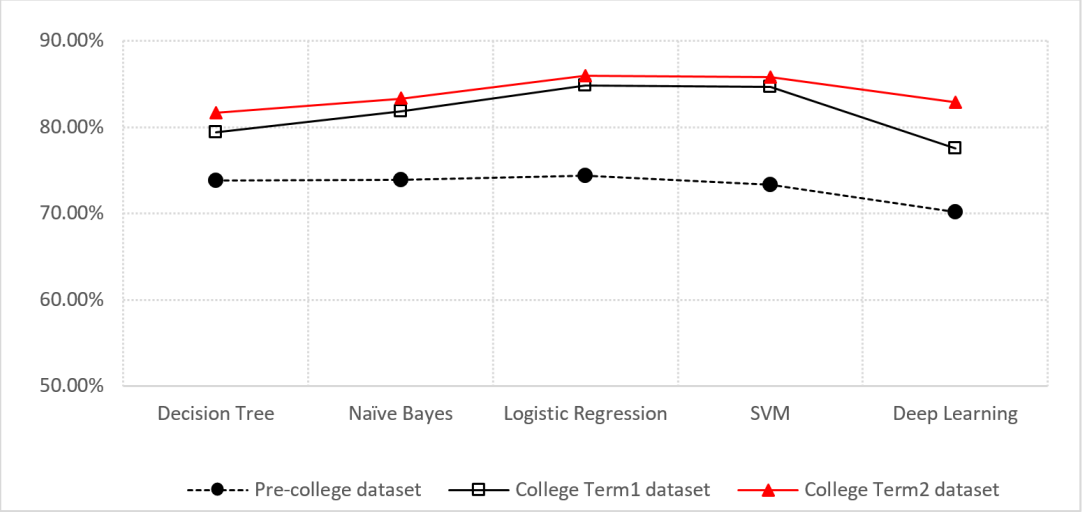


Figure 4.10: AUC performance of standard algorithms on all datasets

The predictive capabilities of all the standard algorithms increases when the College Term 1 and College Term 2 datasets are used, thus making them more effective at predicting dropouts. The results indicate that although

pre-college data can provide a good initial prediction of students likely to dropout, the Term 1 and Term 2 performance data can produce more effective and accurate predictions.

Though not very effective, all standard classifiers can predict dropouts using only pre-college dataset with an AUC score ranging from 70% to 73.8%. Addition of term 1 data enhances the performance of all classifiers, increasing the AUC score by nearly 5% to 11%. The logistic regression and SVM classifiers have shown the most increase of up to 11% AUC while the decision trees' performance has improved by only 5.6%.

A pairwise t-test of significance to test the difference in performance of the Pre-college and College Term 1 dataset reveals that the increase in performance is indeed significant ($\alpha \leq 0.001$, for each algorithm). The difference in performance between the College Term 1 and College Term 2 data set is not very large. However, a pair-wise t-test shows that this increase is also significant and not due to chance, thereby indicating that College Term 2 dataset can provide even more accurate predictions.

4.2.2 Performance by dataset – Ensemble Algorithms

Figure 4.11 shows the AUC score of all ensemble algorithms across the three datasets.

Similar to the behavior of the standard algorithms, the use of college term 1 and term 2 datasets increases the performance of the ensemble

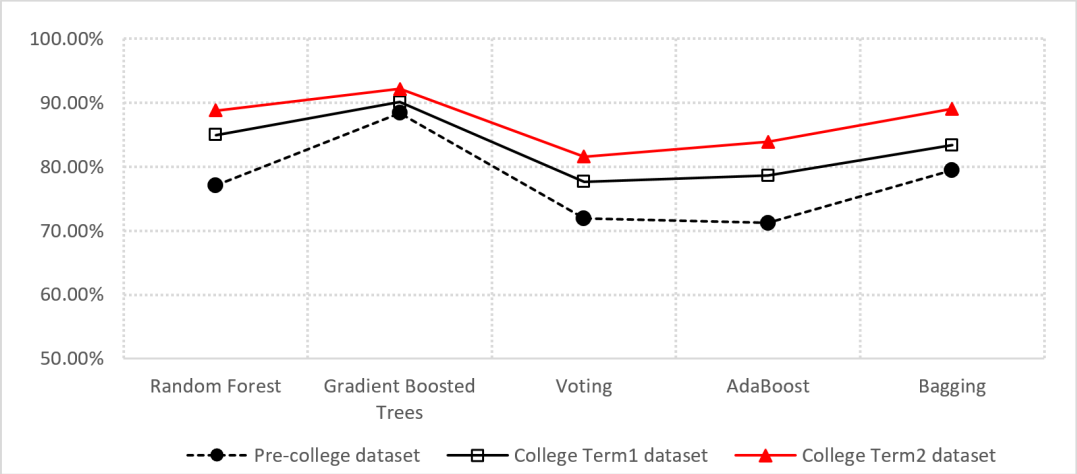


Figure 4.11: AUC performance of ensemble algorithms on all datasets

algorithms as compared to the pre-college. However, the increase is not very high ranging from 2% to 7% for the College Term 1 dataset and 2% to 6% for ensemble algorithms. The gradient boosted trees classifier has shown very little increase proving to be a robust and reliable predictor for all three datasets.

Our results show that potential dropouts and graduates can be predicted as early as on enrollment with an accuracy of 79.31% and AUC of 88.4% with the use of Gradient Boosted Trees algorithm.

4.3 Research Question 3

Which attributes are the top predictors of retention?

To answer this research question, we first analyze the attributes by feature weights to study its relevance with respect to the class label. We begin by calculating the feature weight of each attribute using information gain,

gain ratio, gini index, correlation and chi-squared statistic as explained in Chapter 3. Figure 4.12 shows a comparison of all the feature rankings.

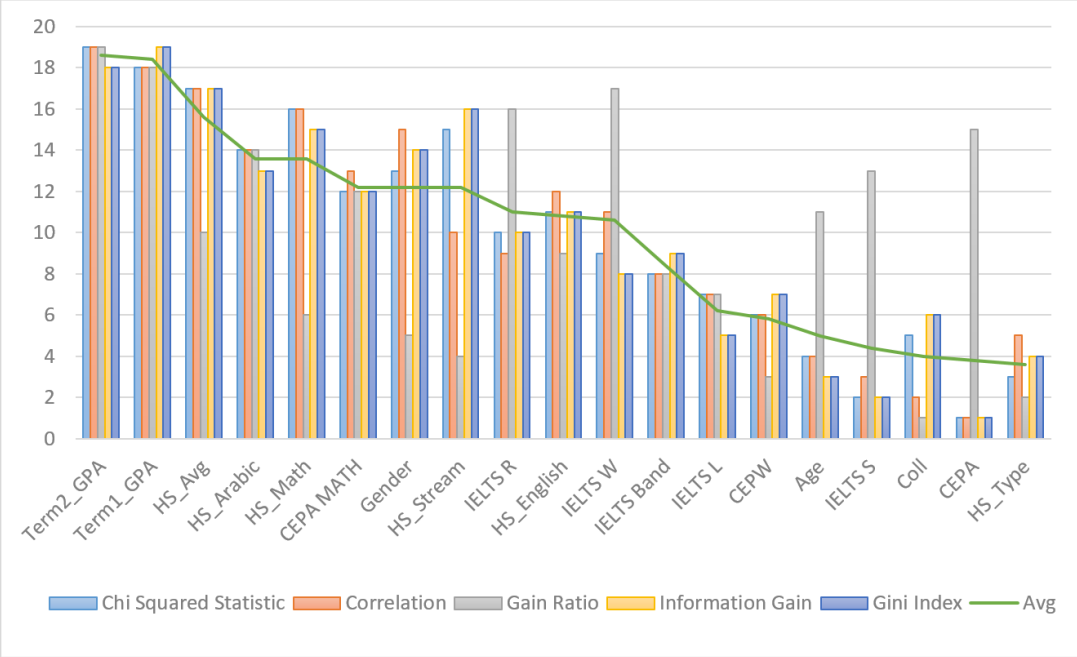


Figure 4.12: Feature Weight

Term1 GPA and Term2 GPA are consistently picked as the most significant predictors by all feature weight algorithms, followed by High School average and then High School Math. Some of the least relevant attributes are Age, IELTS Speaking score, College, CEPA and HS Type (High School Type) which are ranked low by most of the feature weight algorithms.

We also use the decision tree algorithm for its interpretability and ability to identify the top predictor of retention. The root node of the decision tree model shows the most highly influential attribute in classifying dropouts. We also identify the top predictive attributes using the feature weights assigned by the SVM, Logistic regression, Gradient Boosted Trees and Random Forest algorithm across each dataset. A high weight indicates a

higher relevance of the attribute to the prediction.

4.3.1 Pre-college Dataset

Figure 4.13 shows an extract of the decision tree produced with the pre-college dataset. The model reveals that, the high school average, high school stream and IELTS band are the top predictors of retention when the pre-college data set is used.

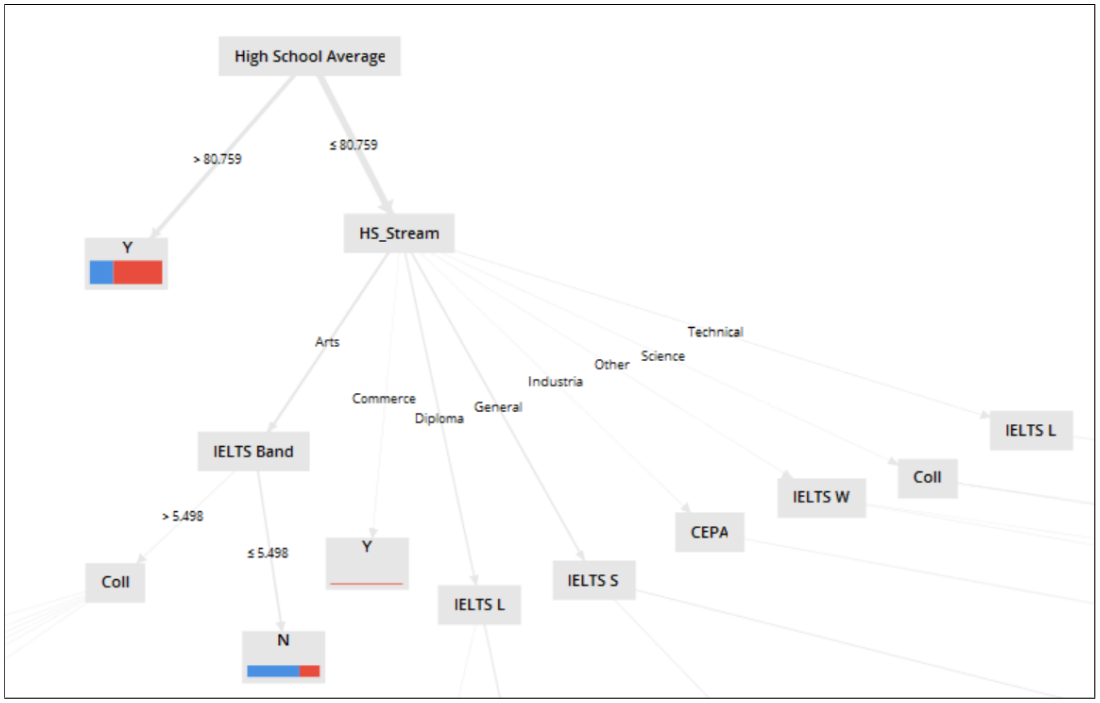


Figure 4.13: Decision Tree model of the pre-college dataset

The SVM algorithm assigns the highest weight to Gender , High School Average, HS Stream and Age. The most relevant attributes identified by the Logistic Regression algorithm is HS Stream , Gender, High School Average, HS Stream and Age.

GBT assigns the highest weight to IELTS Writing, High School Average, Age and IELTS Band, HS Stream, while the top predictors of retention identified by the Random Forest algorithm are High School Math, High School Average, Age, High School Arabic and High School English.

Figure 4.14 shows the overall top predictors of retention using a word cloud to represent the most relevant features by size of the word.

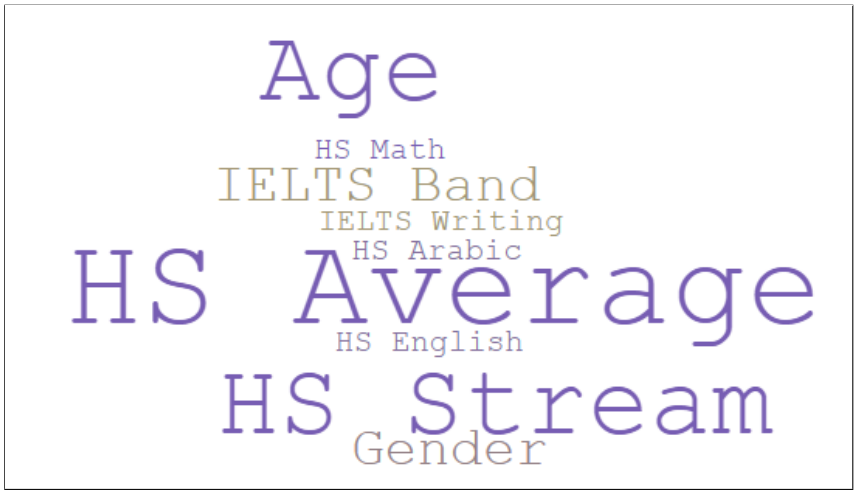


Figure 4.14: Top predictors of retention using the pre-college dataset

High School Average, High School Stream and Age are the top predictors of retention for the pre-college dataset.

4.3.2 College Term 1 Dataset

The decision tree model shows that the Term 1 GPA and the students program are the major predictors of retention when the Term 1 Dataset is used. This indicates that choosing the right program of study is crucial to a student's success in college. Figure 4.15 shows an extract of the decision tree model produced with the term 1 dataset.

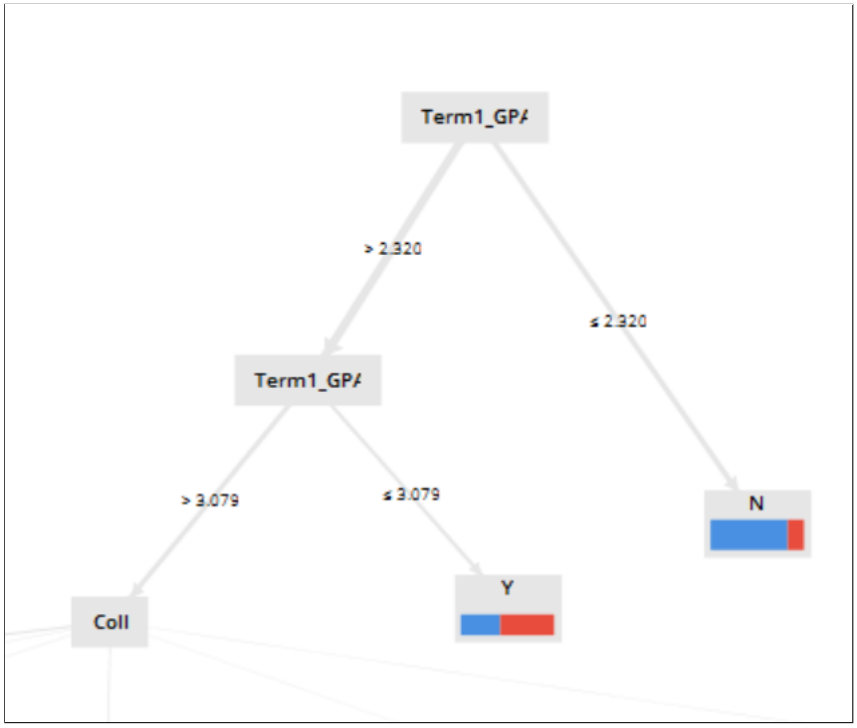


Figure 4.15: Decision Tree model of the college term 1 dataset

SVM and the Logistic Regression algorithm assigns the highest weights to Term1 GPA, HS Stream, and High School Math. Additionally SVM algorithm also picks the program of study as a top predictor.

The GBT algorithm define the top predictors as Term1 GPA, IELTS Writing, IELTS Band, and program of study. Whereas the random forest focuses on Term 1 GPA as well as the high school scores in English and Arabic. All the algorithms consistently pick Term 1 GPA and program of study as the top predictors of retention.

It is interesting to notice that although age is considered as an important predictor in the pre-college dataset, it is not picked as a top predictor by many algorithms in the College Term 1 dataset.

Figure 4.16 shows that the overall top predictors of retention using Col-

lege Term 1 Dataset are the Term 1 GPA and the program of Study.

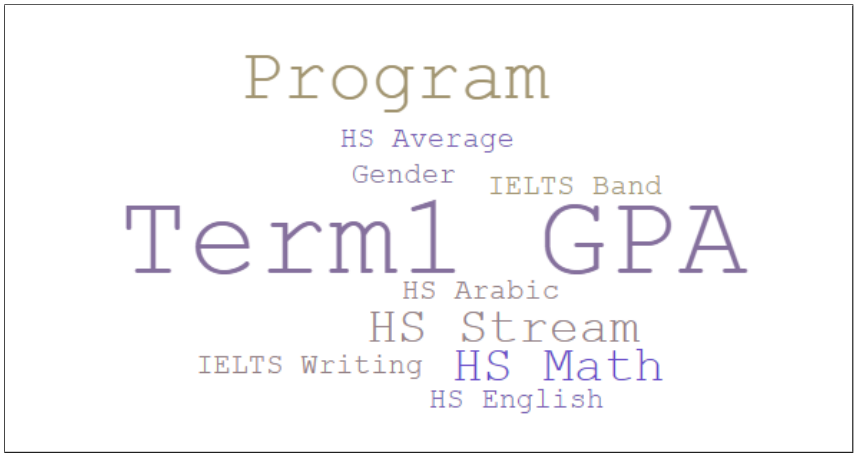


Figure 4.16: Top predictors of retention using the college term 1 dataset

4.3.3 College Term 2

Term 2 GPA, Term 1 GPA , IELTs Band and High School Math scores and are the top predictors of retention identified by the decision tree model using the College Term 2 dataset. Figure shows an extract of the decision tree for the college term 2 dataset.

The SVM , Logistic regression, GBT and RF algorithm consistently assign the highest weight to Term2 GPA, Term1 GPA attributes. The SVM and Logistic Regression algorithm also pick the HS Stream attribute, while the GBT identify the IELTS Band as the top predictor.

The Random Forest algorithm on the other hand assigns a higher weight to high school average and age, thus identifying them as the top predictors of retention.

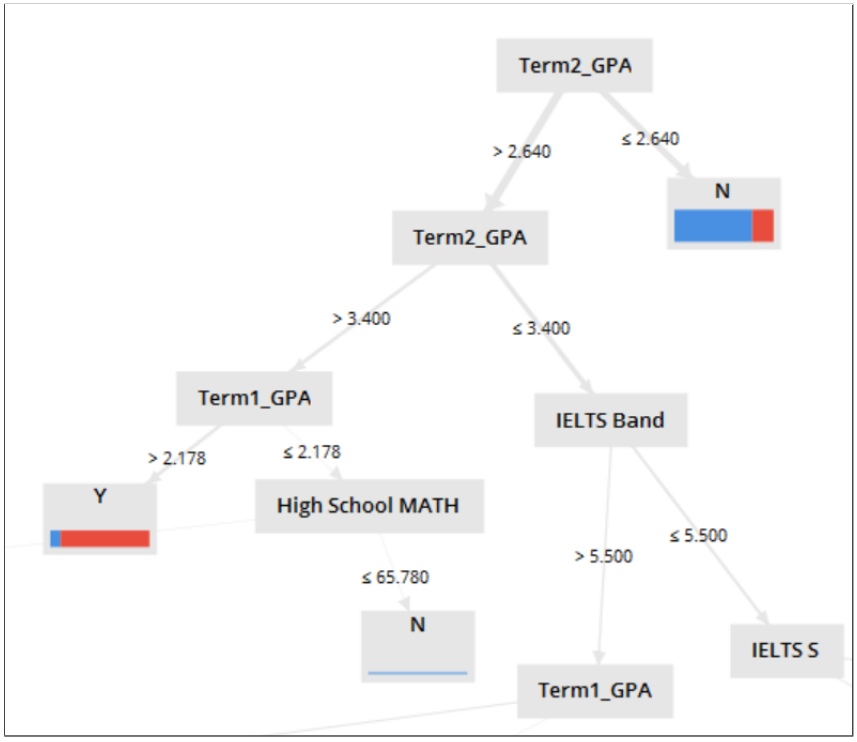


Figure 4.17: Decision Tree model of the college term 1 dataset

Figure 4.18 shows the main predictors of retention using College Term 2 Dataset are the performance in the first and second term.

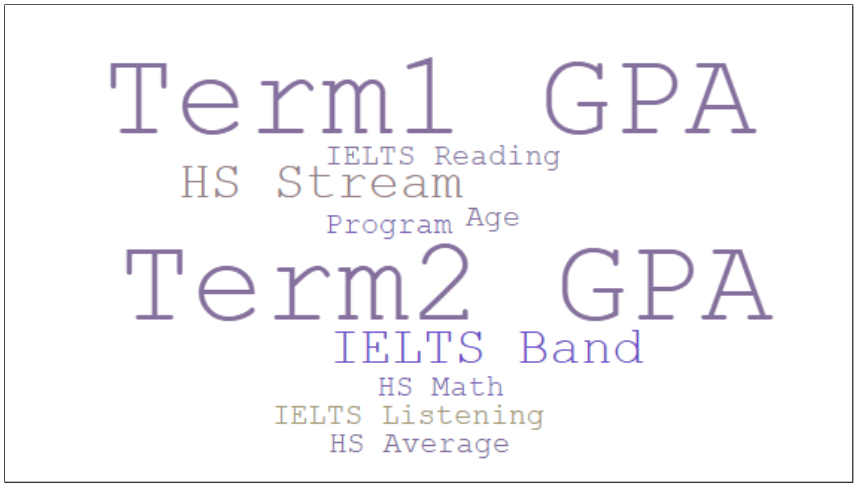


Figure 4.18: Top predictors of retention using the college term 1 dataset

Overall, our results show that college performance is the top predictor of retention during first year of college, while high school average, high school stream and students age is the top predictor in the pre-college dataset.

Chapter 5

Conclusion

The process of steering students from enrollment to graduation is a challenging task that is addressed in this research using a data driven approach. This research fills a practical gap in UAE higher education institutions by building advanced predictive models for an early detection of dropouts. It provides an opportunity for decision makers to leverage new knowledge about students who are at-risk of dropping out, and to implement preemptive measures to improve graduation rates. Our research can also be used to further study retention in UAE based higher education systems.

This research is based on a HEI located in the UAE that offers degree programs to UAE nationals. Therefore, the students of the institution form a homogeneous group belonging to the same nationality, culture and heritage.

The purpose of our research is to predict undergraduate students who are at risk of dropping out without earning a degree. In addition, we seek

to determine the earliest stage when an effective prediction is possible and identify the top predictive factors of retention at each stage.

We apply the structured CRISP-Data Mining methodology to predict student dropout using a dataset of 4,056 student records. The dataset includes student demographic data such as age and gender as well as performance data prior to enrollment (represented by as High School, IELTS and entrance exam scores), and the academic performance data in the first-year of studies.

Our research utilizes ten machine learning algorithms, including five individual algorithms and five ensemble predictors, to build models that can classify a student as a successful graduate or dropout. In addition, models are built by dividing the dataset into three sub datasets of student performance at various stages of the academic journey.

We pre-process and balance the dataset before generating the classification models. The top performing models are compared to determine if the improved performance is statistically significant. In addition, we also identify the top predictors of retention by interpreting the decision tree model and by examining the feature weights assigned by the machine-learning algorithms.

Three research questions are raised in our study. Here we concluded our paper with the answers to those questions.

Question 1: Can machine-learning algorithms effectively predict retention/dropouts in a homogeneous group of students?

We set a threshold requirement of AUC 75% to determine the effectiveness of the standard and ensemble machine-learning algorithms.

Overall, the ensemble algorithms have performed better than standard algorithms across all the three datasets, thus proving to be more reliable and better at handling misclassifications. Among these, the Gradient Boosted Trees is the most effective algorithm performing equally well on all datasets, proving to be a robust and reliable predictor. It achieves an AUC score of 88.4% for the pre-college dataset, 90.1% for College Term1 dataset and 92.2% for College Term 2 dataset.

None of the standard classiffiers are able to meet our threshold performance when the pre-college dataset is used. However, when the dataset is enhanced with College Term 1 performance, all the standard algorithms perform effectively with an AUC score ranging from 77% to 84.8%. Among these the Logistic Regression and SVM performed the best achieving accuracies of up to 77.3%.

Question 2: How early can we predict potential dropouts using machine-learning?

Our research predicts dropouts at a very early stage, using pre-enrollment data, with an accuracy of 79.31% and AUC of 88.40% using the Gradient

Boosted Trees algorithm. Our results will enable the HEI to start remedial support from the first term onwards by directing resources to where they are required the most.

The standard algorithms however, provide only a satisfactory initial prediction of students likely to dropout using the pre-college dataset. Addition of Term 1 and Term 2 performance data produces better predictions with the Logistic Regression and SVM classifiers.

Question 3: Which attributes are the top predictors of retention?

In our research, High School average and IELTS Band are revealed to be the top predictors of retention when a student joins the college. This indicates that students who do well in high school and have a good level of English have a better chance of meeting the academic demands of college.

Interestingly, when the College Term 1 Dataset is used, the pre-enrollment features do not play an important role in predicting dropouts. It is the Term 1 GPA and Program of Study that are the top predictors of graduation. The results reveal that students who perform poorly in their first semester with a GPA below 2.3 are likely to drop out. Hence, the institution should pay close attention to these students and not just to those students who are on probation (GPA < 2.0). Intervention with continuous remedial support at this stage could steer a student to graduation.

The results also show that if students do not choose their program of study wisely, it is likely they will eventually discontinue their studies. This

concurs with the findings of Kirst and Venezia (2004) and Smith and Wertlieb (2005) who state that students often choose their discipline of study based on their interest or prospective career choices without aligning it with their academic capabilities. Hence, this often leads to academic struggle and abandonment of the studies. It is, therefore, essential for the HEI to advise students in wisely choosing their programs of study to ensure success.

In this research, we also discovered that discretizing numerical attributes using binning did not affect the performance of the Decision Tree algorithm significantly. Its predictive capability is equally good and even better for the pre-college dataset and when the numerical values are split based on the decision trees split criterion.

Furthermore, we also discover that k-NN algorithm is unreliable when the SMOTE algorithm is used to balance the dataset. This is because SMOTE technique generates synthetic observations using k-NN algorithm and therefore. It is therefore not surprising that the TNR with k-NN is very high resulting in a biased performance.

It is crucial to collect relevant data to enhance the accuracy of the predictions. While our research has achieved a good accuracy using demographic and performance data, we believe that augmenting the dataset with more detailed data (such as attendance, sponsorship status, and working status) can provide better predictions. Although such data is recorded in the

system, it was unavailable at the time of this study. Another interesting research avenue to pursue would be to expand the study to include the other campuses of the same college and other similar colleges within the region.

The results of our study show that choice of program in the first year is a top predictor of student graduation. An interesting research to pursue would be to investigate recommender systems that effectively recommend degree programs to students to enhance their success rate and steer them to graduation. Our results also show that first term performance is crucial in predicting graduations. Another research avenue to pursue would be to increase the granularity of this study by examining the course work grades for first term courses to predict students who are likely achieve a term 1 GPA below 2.3.

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): 15. 24, 29, 56

ACT (2018). National collegiate retention and persistence to degree rates. 1

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A. and Goodrich, V. (2014). Engagement vs performance: using electronic portfolios to predict first semester engineering student retention, *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, ACM, pp. 103–112. 15, 20, 24, 29, 44, 53

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B. and Addison, K. L. (2015). Who, when, and why: a machine learning approach to prioritizing students at risk of not graduating high school on time, *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, ACM, pp. 93–102. 16, 29

Araque, F., Roldán, C. and Salguero, A. (2009). Factors influencing uni-

versity drop out rates, *Computers & Education* **53**(3): 563–574. 13, 21, 29

Asif, R., Merceron, A., Ali, S. A. and Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining, *Computers & Education* **113**: 177–194. 18, 25, 29

Astin, A. W. (1985). *Achieving educational excellence*, Jossey-Bass,. 11, 12

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016). Predicting student dropout in higher education, *arXiv preprint arXiv:1606.06364* . 17, 27, 29

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions, *JEDM— Journal of Educational Data Mining* **1**(1): 3–17. 2, 14

Bakhshinategh, B., Zaiane, O. R., ElAtia, S. and Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years, *Education and Information Technologies* **23**(1): 537–553. 14

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T. and Popelinsky, L. (2012). Predicting drop-out from social behaviour of students., *International Educational Data Mining Society* . 20, 25, 27, 29, 56

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test

of a causal model of student attrition, *Research in higher education* **12**(2): 155–187. 11, 12

Berger, J. B., Ramírez, G. B. and Lyons, S. (2005). Past to present, *College student retention: Formula for student success* **1**. 11

Bowers, A. J., Sprott, R. and Taff, S. A. (2012). Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity, *The High School Journal* pp. 77–100. 57

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C. and Lykeridou, K. (2015). Examining students graduation issues using data mining techniques-the case of tei of athens, *AIP Conference Proceedings*, Vol. 1644, AIP, pp. 255–262. 1, 4

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357. 26, 27, 47

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F. and Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, *Computers in Human Behavior* **73**: 247–256. 15, 21, 27, 30, 47

Danilowicz-Gösele, K., Lerche, K., Meya, J. and Schwager, R. (2017).

Determinants of students' success at university, *Education economics* **25**(5): 513–532. 21, 30

Dekker, G. W., Pechenizkiy, M. and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study., *International Working Group on Educational Data Mining* . 16, 22, 27, 30, 42, 56

Delen, D. (2011). Predicting student attrition with data mining methods, *Journal of College Student Retention: Research, Theory & Practice* **13**(1): 17–35. 2, 15, 19, 20, 24, 27, 30, 42, 53, 56

Demetriou, C. and Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism: Retention theories past, present and future, *Proceedings of the 7th National Symposium on student retention*, pp. 300–312. 1

Djulovic, A. and Li, D. (2013). Towards freshman retention prediction: a comparative study, *International Journal of Information and Education Technology* **3**(5): 494–500. 17, 30

Dutt, A., Ismail, M. A. and Herawan, T. (2017). A systematic review on educational data mining, *IEEE Access* **5**: 15991–16005. 14

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil, *Journal of Business Research* **94**: 335–343. 21, 30

Gershenfeld, S., Ward Hood, D. and Zhan, M. (2016). The role of first-semester gpa in predicting graduation rates of underrepresented students, *Journal of College Student Retention: Research, Theory & Practice* **17**(4): 469–488. 13, 21, 30

Guarín, C. E. L., Guzmán, E. L. and González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining, *IEEE Revista Iberoamericana de tecnologias del Aprendizaje* **10**(3): 119–125. 6, 22, 25, 27, 30, 47, 56

GulfNews (2017). New ratings system for uae universities education quality.

URL: *https://www.khaleejtimes.com/nation/new-ratings-system-for-uae-universities-education-quality* 1, 4

Hoffait, A.-S. and Schyns, M. (2017). Early detection of university students with potential difficulties, *Decision Support Systems* **101**: 1–11. 2, 4, 6, 23, 24, 31, 56

Huang, S. and Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models, *Computers & Education* **61**: 133–145. 15, 31, 56

Jacob, J., Jha, K., Kotak, P. and Puthran, S. (2015). Educational data mining techniques and their applications, *Green Computing and Inter-*

net of Things (ICGCIoT), 2015 International Conference on, IEEE, pp. 1344–1348. 2

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R. and Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative, *Journal of Learning Analytics* **1**(1): 6–47. 15, 19, 21, 27, 31

Kirst, M. and Venezia, A. (2004). *From high school to college: Improving opportunities for success*, San Francisco: Jossey-Bass. 85

Kovacic, Z. (2012). Predicting student success by mining enrolment data. 6, 15, 19, 20, 21, 25, 31, 56

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*, Vol. 26, Springer. 54, 55

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R. and Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes, *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, pp. 1909–1918. 16, 31, 53

Larose, D. T. and Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*, John Wiley & Sons. 2

Larsen, M. S., Kornbeck, K. P., Kristensen, R. M., Larsen, M. R. and

Sommersel, H. B. (2012). Dropout phenomena at universities: What is dropout? why does, *Education* **45**: 1111–1120. 13

Levitz, R. S., Noel, L. and Richter, B. J. (1999). Strategic moves for retention success, *New directions for higher education* **1999**(108): 31–49. 2, 5

Ling, C. and Sheng, V. (n.d.). Cost-sensitive learning and the class imbalance problem. 2011, *Encyclopedia of Machine Learning: Springer* . 27

Ma, J., Pender, M. and Welch, M. (2016). Education pays 2016: The benefits of higher education for individuals and society. trends in higher education series., *College Board* . 1

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H. and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students, *Expert Systems* **33**(1): 107–124. 15, 20, 23, 24, 27, 32, 47

Márquez-Vera, C., Morales, C. R. and Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* **8**(1): 7–14. 15, 20, 23, 24, 27, 31

Martinho, V. R., Nunes, C. and Minussi, C. R. (2013). Prediction of school dropout risk group using neural network, *Computer science and informa-*

tion systems (FedCSIS), 2013 federated conference on, IEEE, pp. 111–114. 19, 32

Mayra, A. and Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in ecuador, *Global Engineering Education Conference (EDUCON), 2018 IEEE*, IEEE, pp. 1238–1242. 5, 23

Miguéis, V. L., Freitas, A., Garcia, P. J. and Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach, *Decision Support Systems* **115**: 36–51. 18, 32, 53

Mirowsky, J. and Ross, C. E. (2005). Education, cumulative advantage, and health, *Ageing International* **30**(1): 27–62. 5

Natek, S. and Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions, *Expert systems with applications* **41**(14): 6400–6407. 32

NCES (2018). Undergraduate retention and graduation rates.
URL: *https://nces.ed.gov/programs/coe/indicator*$_c$*tr.asp*4

NSCResearchCenter (2019). Persistence retention – 2018.
URL: *https://nscresearchcenter.org/snapshotreport33-first-year-persistence-and-retention/* 4

Oztekin, A. (2016). A hybrid data analytic approach to predict college graduation status and its determinative factors, *Industrial Management & Data Systems* **116**(8): 1678–1699. 20, 22

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works, *Expert systems with applications* **41**(4): 1432–1462. 14

Perez, B., Castellanos, C. and Correal, D. (2018). Applying data mining techniques to predict student dropout: A case study, *2018 IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI)*, IEEE, pp. 1–6. 17, 20, 25, 32

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making, *Big data* **1**(1): 51–59. 53

Raju, D. and Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models, *Journal of College Student Retention: Research, Theory & Practice* **16**(4): 563–591. 6, 17, 18, 22, 32

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005, *Expert systems with applications* **33**(1): 135–146. 14

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): 601–618. 13, 14

Romero, C. and Ventura, S. (2013). Data mining in education, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **3**(1): 12–27. 2, 38

Rubiano, S. M. M. and Garcia, J. A. D. (2015). Formulation of a predictive model for academic performance based on students' academic and demographic data, *2015 IEEE Frontiers in Education Conference (FIE)*, IEEE, pp. 1–7. 22, 32

Seidman, A. (2005). Minority student retention: Resources for practitioners, *New directions for institutional research* **2005**(125): 7–24. 1

Seni, G. and Elder, J. F. (2010). Ensemble methods in data mining: improving accuracy through combining predictions, *Synthesis Lectures on Data Mining and Knowledge Discovery* **2**(1): 1–126. 53

Shahiri, A. M., Husain, W. et al. (2015). A review on predicting student's performance using data mining techniques, *Procedia Computer Science* **72**: 414–422. 22

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining, *Journal of data warehousing* **5**(4): 13–22. 9, 34

Smith, J. S. and Wertlieb, E. C. (2005). Do first-year college students' expectations align with their first-year experiences?, *NASPA journal* **42**(2): 153–174. 85

Stinebrickner, R. and Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model, *Journal of Labor Economics* **32**(3): 601–644. 21

Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M. and Appleton, J. (2014). Predicting student risks through longitudinal analysis, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1544–1552. 6, 19, 22

Thammasiri, D., Delen, D., Meesad, P. and Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition, *Expert Systems with Applications* **41**(2): 321–330. 15, 26, 27, 47

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research, *Review of educational research* **45**(1): 89–125. 2, 11, 12, 20, 22, 24

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G. and Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction, *Simulation Modelling Practice and Theory* **55**: 1–9. 2

Witten, I. H., Frank, E. and Mark, A. (2011). Hall. 2011, *Data mining: Practical machine learning tools and techniques* **3**. 49, 51, 52

Yu, C. H., DiGangi, S., Jannasch-Pennell, A. and Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year, *Journal of Data Science* **8**(2): 307–325. 1, 20, 24, 33

Yukselturk, E., Ozekes, S. and Türel, Y. K. (2014). Predicting dropout stu-

dent: an application of data mining methods in an online education program, *European Journal of Open, Distance and E-learning* **17**(1): 118–133. 15, 21, 23, 24, 28, 33

Zhang, Y., Oussena, S., Clark, T. and Kim, H. (2010). Using data mining to improve student retention in higher education: a case study, *In International Conerence on Enterprise Information Systems*, Citeseer. 20, 27, 33

Zollanvari, A., Kizilirmak, R. C., Kho, Y. H. and Hernández-Torrano, D. (2017). Predicting students' gpa and developing intervention strategies based on self-regulatory learning behaviors, *IEEE Access* **5**: 23792–23802. 24, 33