



**Predicting Student Withdrawal from
UAE CHEDS Repository using Data Mining
Methodology**

توقع الطالب المنسحب من مستودع البيانات CHEDS التابع لدولة الامارات
باستخدام منهجية التنقيب عن البيانات

by

AHMAD ABDULLA BINEID

**Dissertation submitted in partial fulfilment
of the requirements for the degree of
MSc INFORMATION TECHNOLOGY MANAGEMENT
at
The British University in Dubai**

November 2022

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that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.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.



Signature of the student

COPYRIGHT AND INFORMATION TO USERS

The author whose copyright is declared on the title page of the work has granted to the British University in Dubai the right to lend his/her research work to users of its library and to make partial or single copies for educational and research use.

The author has also granted permission to the University to keep or make a digital copy for similar use and for the purpose of preservation of the work digitally.

Multiple copying of this work for scholarly purposes may be granted by either the author, the Registrar or the Dean of Education only.

Copying for financial gain shall only be allowed with the author's express permission.

Any use of this work in whole or in part shall respect the moral rights of the author to be acknowledged and to reflect in good faith and without detriment the meaning of the content, and the original authorship.

ABSTRACT

Early prediction of a student who is at risk of course dropout leads to student retention in the study course. The percentage of student dropout in higher education sector is high, and affects the students' careers negatively and the institute's program continuation.

The purpose of this study is to predict and identify students who are likely to withdraw from an institute. This identification assists the institute's advisor to take precautionary measures to retain this group of students. Also, the study aims to find the variable that is most efficient to lead to student dropout prediction.

To fulfil the study's aim, CRISP method was followed after reviewing research papers. A dataset of 1272 students' data in size from Central Higher Education Data Store (CHEDS) has been fetched from Dubai's governmental higher education institute. The demography of students is international background. Several model classifiers from Standard and ensemble were implemented to find the best answer to the research questions. Receiver Operator Characteristic (ROC) based on Area Under Curve (AUC) was used to assess the result plus other metrics.

Research outcome, results showed that students who had low GPA, average register credit hours and fluctuating student's enrollment status were more likely to withdraw from study course. Random Forest classifiers demonstrated the highest performance in prediction, and scored 87.8% in AUC with an accuracy of 84.82%. GPA and average register credit hours attributes were the most effective contributor in prediction.

نبذة موجزة

يعد التنبؤ بالطلاب المعرضين لخطر ترك مقاعد الدراسة من أهم الخطوات التي تؤدي إلى استمرارية الطلاب بالبرامج الدراسية، والحصول على درجات علمية. حيث تعتبر نسبة تسرب الطلاب من قطاع التعليم العالي مرتفعة؛ ولها تأثير سلبي على قطاع التعليم واستمرارية البرامج الأكاديمية ومسيرة الطلاب العملية.

الغرض من هذه الدراسة هو التنبؤ وتحديد الطلاب المحتمل عدم إكمالهم البرنامج الأكاديمي. حيث يخدم هذا النوع من التنبؤ المرشد الأكاديمي للطالب في وضع الإجراءات الوقائية للاحتفاظ بهذه المجموعة من الطلاب. وتهدف الدراسة أيضاً إلى تحديد المتغير الأكثر فاعلية للتنبؤ بانسحاب الطالب.

لتحقيق هدف الدراسة، تم اتباع منهجية CRISP بعد مراجعة الأوراق البحثية بنفس المجال. استُخدم بالبحث مجموعة بيانات ترجع إلى 1272 طالب من مخزن بيانات التعليم العالي المركزي (CHEDS) من كلية للتعليم العالي التابعة لحكومة دبي. خلفية التركيبة السكانية للطلاب هي خلفية دولية. تم تنفيذ العديد من المصنفات النموذجية من المصنف القياسي والمصنف المجموعي للعثور على أفضل إجابة لأسئلة البحث. لتقييم نتائج البحث، استخدمت خاصية مشغل المستقبل (ROC) على أساس المساحة تحت المنحنى (AUC) بالإضافة إلى مقاييس أخرى.

تم التوصل من خلال البحث أن الطالب الأكثر عرضة لانسحاب من البرنامج الأكاديمي هو الذي لديه معدل تراكمي منخفض، ومتوسط ساعات التسجيل بالساعات الدراسية المعتمدة، وحالة متقلبة بالتسجيل بالفصول الدراسية المتوالية. تُظهر مصنفات Random Forest المجموعي أعلى أداء في التنبؤ، حيث سجلت AUC ما نسبته 87.8% وبدقة تبلغ 84.82%. تسهم بيانات المعدل التراكمي (GPA attribute) وبيانات متوسط الساعات المعتمدة للتسجيل الدراسي (hours attributes average register credit) من أكثر العوامل المؤثرة في التنبؤ.

ACKNOWLEDGEMENTS

الحمد لله

I thank **Allah**, lord of majesty, bounty, favour, benevolence, for his assistance in completing my scientific research, in which I hope to benefit and contribute to science and education. Without his help, I would not have been able to complete it.

And I dedicate this research to the one who encouraged us to seek knowledge and learning, to the **Prophet Muhammad**, may Allah's prayers and peace be upon him. He said in the meaning of a Hadith: "Seeking knowledge is an obligation upon every Muslim".

I would also like to thank the one who guided and advised me to the right approach in my research, **Professor Dr. Sherief Abdallah**, for his time and effort in completing this thesis to the fullest.

In addition, a special dedication to my father, and especially to my mother, who left us before I had the honor of her look at this research. I thank and appreciate them very much for what they scarified in my upbringing and education and for urging me to seek knowledge.

In conclusion, I extend my warmest regards to my family and children for their support and encouragement to accomplish this scientific work. I will not forget to thank and include the endless list of my friends for their encouragement to carry out this research.

TABLE OF CONTENTS

CONTENTS.....	I
LIST OF FIGURES	III
LIST OF TABLES	IV
INTRODUCTION.....	1
1.1 RESEARCH MOTIVATION	1
1.2 RESEARCH OBJECTIVE.....	2
1.3 RESEARCH METHODOLOGY.....	3
1.4 RESEARCH STRUCTURE	3
LITERATURE REVIEW.....	4
2.1 THEORETICAL MODEL OF STUDENT ATTRITION	4
2.2 EDM APPROACHES TO STUDY STUDENT ATTRITION.....	5
2.2.1 <i>Introduction to data mining and EDM</i>	5
2.2.2 <i>EDM studies' approaches</i>	7
2.3 FACTORS THAT CONTRIBUTE TO STUDENT ATTRITION	8
2.3.1 <i>Student's demographics, social and psychological factors</i>	9
2.3.2 <i>Student's prior performance and academic factors</i>	10
2.3.3 <i>Student's engagement and institutional factors</i>	11
2.3.4 <i>Financial factors</i>	12
METHODOLOGY.....	15
3.1 BUSINESS UNDERSTANDING	16
3.2 DATA UNDERSTANDING	17
3.2.1 <i>Data Acquisition</i>	18
3.2.2 <i>CHEDS Database</i>	18
3.3 DATA PREPARATION.....	20
3.3.1 <i>Data Cleaning</i>	20
3.3.2 <i>Data Integration</i>	24
3.3.3 <i>Data Transformation</i>	25
23.4 DATA MODELING	32
3.4.1 <i>Modeling Technique</i>	32
3.4.2 <i>Testing Design</i>	35
3.4.3 <i>Building Model</i>	35
3.5 MODEL ASSESSMENT AND EVALUATION	39
3.6 DATA DEPLOYMENT	41
RESULT AND DISCUSSION.....	42

4.1 RESEARCH QUESTION 1 AND 2.....	42
4.1.1 <i>Best Performance Model</i>	42
4.1.2 <i>Least Performance Model</i>	47
4.1.3 <i>Summary of Planned Models</i>	48
4.2 RESEARCH QUESTION 3.....	51
CONCLUSION.....	57
FINDING AND CONTRIBUTION.....	57
LIMITATION.....	58
FUTURE WORK.....	59
REFERENCES.....	60

List of Figures

FIGURE 1: CRISP-DM WORKFLOW	16
FIGURE 2: IMPUTE MISSING VALUE OPERATOR	27
FIGURE 3: ONE HOT ENCODING OPERATOR	28
FIGURE 4: NORMALIZATION OPERATOR USED	29
FIGURE 5: FEATURE SELECTION OPERATORS	30
FIGURE 6: TOTAL DROP PER GENDER.....	30
FIGURE 7: DROP BASED ON CITY WISE	31
FIGURE 8: PERCENTAGE OF DROP PER CITY	31
FIGURE 9: NATIONALITY DROP PERCENTAGE	32
FIGURE 10: ACCURACY AND PRECISION DIFFERENCE	40
FIGURE 11: ROC VARIANT BASED ON AUC VALUE	41
FIGURE 12:DECISION TREE AUC	44
FIGURE 13: ILLUSTRATION OF DT AND WITH DISCRETIZE	45
FIGURE 14: SUMMARY OF AUC DIAGRAM.....	50
FIGURE 15: TOP FIVE FEATURE WEIGHT ATTRIBUTES.....	51
FIGURE 16: ATTRIBUTE CORRELATION MATRIX	52
FIGURE 17: DECISION TREE MODEL.....	56

List of Tables

TABLE 1: STUDIED RESEARCH PAPER SUMMARY	14
TABLE 2: COMPARISON OF ENROLLMENT AND GRADUATE BASED ON GPA.....	19
TABLE 3: CHEDS REPORT AVAILABILITY IN STUDY	20
TABLE 4: ATTRIBUTES CHANGES IN PROGRESSINVE YEARS	22
TABLE 5: NUMBER OF CHEDS ATTRIBUTES IN ACADEMIC YEAR	22
TABLE 6: SELECTED ATTRIBUTES IN STUDY	23
TABLE 7: SEMESTERS CODES	24
TABLE 8: FINAL SELECTED ATTRIBUTES LIST	24
TABLE 9: MISSING INFORMATION PERCENTAGE PER ATTRIBUTE	27
TABLE 10: ONE HOT ENCODING CONVERSION SAMPLE	28
TABLE 11: ATTRIBUTES AFTER AUTOMATED DATA PROCESSING	29
TABLE 12: DT SET PARAMETERS	37
TABLE 13: K-NN SET PARAMETERS	37
TABLE 14 SVM SET PARAMETERS	37
TABLE 15: DEEP LEARNING SET PARAMETERS	37
TABLE 16: RF SET PARAMETERS.....	38
TABLE 17: BAGGING SET PARAMETERS.....	38
TABLE 18: BOOSTING SET PARAMETERS.....	38
TABLE 19: GBT SET PARAMETERS	39
TABLE 20: DT MODEL PERFORMANCE.....	43
TABLE 21: COMPARISON BETWEEN DT AND WITH DISCRETIZE.....	46
TABLE 22: DT PERFORMANCE OF SPLIT CRITERIONS	46
TABLE 23: BEST AUC RANDOM FOREST PERFORMANCE.....	47
TABLE 24: K VALUE AND ERROR PERCENTAGE.....	47
TABLE 25: COMPARISON OF AUC NB AND KNN.....	48
TABLE 26: SUMMARY ALL CLASSIFIERS' MODELS	49
TABLE 27: TOP VARIABLE WEIGHT CORRELATION.....	52
TABLE 28: INDEPENDENT VARIABLE WEIGHT TO CLASS LABEL.....	55

Chapter 1

Introduction

Retention is a valuable word when it comes to talented people. Retention is used by human resource departments of pioneer business firms to keep talented employees. Retention is used in trading to keep valuable purchasers. In Education, the term “student retention” has become a well-known terminology that is used hand in hand with the students’ withdrawal phenomenon from academic programs.

Retaining a student in a higher education institute and having him to continue to study efficiently is a significant challenge. A student’s withdrawal from a study course is related to many complex factors. Factors vary from Student’s demographic, social factors, psychological factors, financial factors, prior performance, institutional policy & procedure, and many more.

1.1 Research Motivation

Higher Education institutes need new ways to retain students. The classic approach of providing students’ advisors is not enough in maintaining students. Retention rate of students in higher education institutes scores differently from one to another. University of Texas at Austin scored 95% in student retention. On the other hand, Baylor University scored 88%. These details are based on US news and world reports of the year 2016.

The withdrawal percentage is at a higher rate when it comes to middle east high education sector. There is a need to come up with new discoveries and approaches to find out students at risk.

Data mining and knowledge discovery has been used for a long time in business, engineering, and many other fields. Recently, data mining has been used in the education sector under the term EDM, which stands for Education Data Mining. EDM uses mostly CRISP-DM method classification prediction.

Predicting and preventing any student’s dropping out of a study course is crucial to the student’s career, the institute’s effectiveness rank and the country’s education rank among region countries. The interest in the following three elements were the motivation behind conducting this data mining research:

First: Conducting education data mining on a unique and standardized database of the UAE's Ministry of Education (Central Higher Education Data Store) CHEDS database. As far as my knowledge goes and according to research papers in Google scholar; this research is the first from its kind in applying EDM on CHEDS database. Such research implementation would motivate data scientists to build up on this research's outcome and utilize the contribution from EDM's researcher instead of individual institute's database to the standardized CHEDS database. The result of the EDM's model would be applicable to any higher education institute recognized and accredited by MoE. This would give a chance to implement the developed prediction model from this research on multiple institutes; and streamline the efforts of data scientists to improve the prediction model. Also, it would be interesting if MoE could host the algorithm in CHEDS portal; and offer it to researchers and data scientists to be as an open-source prediction model on CHEDS portal with anonymized real dataset.

Second: Participating in improving UAE's HEI's effectiveness by identifying students of withdrawal risk. This identification helps institutes to retain the student in the study. The identification would be applicable to all Higher Education Institutes that follow data statistic's reporting CHEDS template format. For that, the vast range of implementation would lead to improving UAE's Education sector and UAE's index among other countries.

Third: The fact that previous scientific education data mining work commenced in western countries would not be applicable on GCC countries and specially UAE. As the environment does not match with western countries from a social and cultural point of view. Also, the local data mining was implemented on non-standardized dataset with multiple dataset integration of pre-college information, which makes the comparison between studies unequal since attributes vary.

1.2 **Research Objective**

The purpose of this research is to support the higher education sector with predication classification using EDM approach. The objectives are to find students at risk of withdrawal from study course, recognize the dominant variables to predict withdrawing students from the dataset and to estimate the accuracy of the purposed classification prediction model.

The following are the research questions developed to navigate the research phases:

- 1. Which students are likely to withdraw from the institute?**
- 2. How accurate a prediction factor is in making a decision in an organization?**
- 3. Which variables are more efficient at predicting a student's withdrawal?**

1.3 Research Methodology

CRISP-DM (Cross-Industry Standard Process for Data Mining) is often selected from the most popular methodologies of data mining. There is another approach called SEMMA (Sample, Explore, Modify, Model, Assess), but it is not commonly selected; since it is not involved in business understanding and deployment.

CRISP-DM consists of six stages, which are business understanding, data understanding, data preparation, data modeling, data evaluation and data deployment.

In data preparation and data modeling two kinds of software were used. For data preparation MS Excel – Power Query was used and RapidMiner Studio Educational version v9.10.011 was used for modeling.

1.4 Research Structure

This research consists of 5 chapters, starting with an introduction to the study. Chapter 2 presents the literature review of the research papers in education data mining in regarding student withdrawal prediction. In this chapter, a comparison was made on research aim, factors, used algorithms and evaluation approaches. Chapter 3 highlights the method used to develop a classification model. Chapter 4 conveys results presentation and evaluation of outcomes. Finally, Chapter 5 sums up research conclusion and findings.

Chapter 2

Literature Review

Student attrition can be defined as student withdrawal from an academic program or semester for several reasons, such as: official withdrawal from a program or ‘not show’ by not attending classes, failed in course because a student stopped from continuing study, and registration in a program without continuing the enrollment process (Ahmad Tarmizi et al. 2019). Student attrition can be found when the number of graduated students in a semester or year is less than the number of enrollments in the same program (Beer and Lawson 2017). Factors leading to attrition can vary and get complex in non-homogenous students’ population (Ahmad Tarmizi et al. 2019). Non-homogenous students’ population makes it harder for the university to predict the students with a higher probability to withdraw from a study course and the nature of the cause to withdraw (Beer and Lawson 2017).

This study focuses on predicting the student dropout (student attrition) from HEI. Earlier literatures are reviewed based on the work that have been made to predict students with the most probability to withdraw due to lack of performance using EDM methods and techniques.

2.1 Theoretical model of student attrition

Theoretical model of student attrition in education has been found by several scientists and researchers, where they analyzed numerous variables that lead to a student’s decision to drop-out from a course or to discontinue enrollment in the next semesters (Spady 1970; Spady 1971; Tinto 1975; Bean and Metzner 1985).

Spady (Spady 1970) built the models on five independent variables. The most predominant variable is (social integration). These five independent variables are connected to the dependent variable (withdrawal decision); with two intervention variables, which are institutional commitment and satisfaction. But Spady (Spady 1971) revised the model after conducting a test on the model using first year undergraduate information; by adding two important improvements. The improvements are: first, including relationship structure and friendship support, second, re-ordering relationship of model elements.

The second major theoretical model on student attrition has been developed by Tinto (Tinto 1975). Tinto built student attrition module based on Spady (Spady 1971) works, and Tinto's module for university student attrition remains to be the most successful and tested module. Tinto's theory states that the student's tendency to continue studying in college is higher when a student felt connected to college's social and academic life. This is enhanced positively with accumulative interaction of variables such as: student background, student commitment to study in university and interactions between student and faculty. These variables lead to connecting students to social and academic life. Tinto's theory has been examined by other researchers who confirmed its predictive to student attrition (Pascarella and Chapman 1983). Tinto indicated that academic integration has a higher effect on social integration for longer student observation.

Bean and Metzner (Bean and Metzner 1985) came up with a different approach than Spady and Tinto. Their approach of developing theory was based on nontraditional students that are not affected by social variables when deciding to dropout from a study course. Bean and Metzner developed their theory of nontraditional student attrition based on Bean's (Bean 1980) work. The theory assumed that student attrition is caused by four variables: academic performance, intention to leave due to psychological outcomes and academic variables, school graduation performance result and educational goals, and environmental variables which is the primary variable in their theory.

There are two additional compensatory elements that have higher influence on the model. First, academic variable and environmental variable. Bean and Metzner (Bean and Metzner 1985) believed that with low academic values, the nontraditional student could stay in study due to high environmental value. Second, psychological outcomes and academic outcomes. Bean and Metzner believed that nontraditional student with low academic outcomes could continue studying if psychological outcomes are positive (Bean and Metzner 1985).

2.2 EDM approaches to study student attrition

2.2.1 Introduction to data mining and EDM

Data mining methodology is one of the most reliable sciences for evaluating vital information from datasets. Data mining predicts valuable hidden information by using

mining methods, which result in improving conclusions and decisions (Salloum et al. 2017). Data mining is one of the processes used in the knowledge discovery process and it is involved in the extraction of useful information from a data warehouse using systematic methodologies (Azevedo 2019).

Logs, records and dataset of a system consist of valuable historical information recorded in certain patterns which data mining algorithms can show these patterns and trends. These conclusions contribute in enhancing the quality of system working procedures (Azevedo 2019). Functions of data mining in educational processes have leveraged student performance, staff and institute decision (Fernandes et al. 2019). Also, it is a multi-dimensional science consisting of multiple features such as artificial intelligence, statistics, visualization, information technology and more (Alomari, K.M., AlHamad, A.Q. and Salloum 2019).

Educational data mining is the science of extracting valuable data and information from accumulated records and logs in the educational field. It is a data-driven process to diagnose and find student weak points, performance failure issues as well as prediction of future progress (Adekitan and Noma-Osaghae 2019).

Educational data mining's goal is to enhance student learning environment along with institute efficiency (Bucos and Drăgulescu 2018). Educational dataset consists of useful information used to analyze students' results and predict future performance. It presents unseen relations among various learning characteristics, student performance and behaviors (Adekitan and Noma-Osaghae 2019). Results achieved from data mining of educational databases can be used to enhance teaching delivery methods and quality of educational system (Daradoumis et al. 2019). Mined education's data can be used in the development of education fields like improving communication methods, redesigning courses, modifying assessment procedures and more to increase the quality of education (Baepler and Murdoch 2010).

Studying academic performance and predicting students' results are rich fields for machine learning. Machine learning application in education can be found in predicting student performance, predicting graduation vs. drop out potential, assessing learning process, highlighting and identifying learning risk, evaluating students' feedback, evaluating institute administration and many other useful applications (Adekitan and Noma-Osaghae 2019).

Data-driven knowledge is a trend in research field. Researchers are more focused on

educational dataset due to the richness of its databases. Output accuracy depends on the size and quality of a database (Almarabeh 2017). Numerous literatures presented the importance of educational data mining and its outstanding contribution in providing analysis and management support tools to assist institutes in successful decision making.

In such a field where data-driven approach is the method in educational data mining; privacy is a concern that should be respected while finding the commercial value to sustain the business through data mining output (Lynch 2017). The increased application of educational data mining improved the educational systems by stabilizing and integrating students and faculty. The idea of educational data mining development has been appearing in various educational institutes. Examples of areas influenced by these developments are: student performance problems and institute quality effectiveness (Awad et al. 2020).

2.2.2 EDM studies' approaches

There are two main approaches which are used by researchers to analyze the attrition in education. The first approach is survey-based research; the second is analytical based research. The first is based on questionnaire answers from participants, which is not our scope in this study. The second is based on technological development and computing intelligence in analyzing the educational dataset.

Research papers on EDM show a number of factors that lead to academic continuation and success or students' attrition. The purpose of Natek and Zwilling's (Natek and Zwilling 2014) literature is to review and predict the most relevant student attrition's factors; and to focus on the used methods and algorithms that lead to the most successful prediction rate. Natek and Zwilling (Natek and Zwilling 2014) analyzed 30 literatures focusing on education retention and students' attrition, and found six attributes that are most often used. The number one attribute out of the six is cumulative grade points average (CGPA). Similarly, internal assessment attributes scored top of the list. In the second group of the most used attributes are demographic characteristic and external assessment, meaning high school achievement. In the third attributes group are curricular activities and social collaboration.

In a similar study to Natek and Zilling's (Natek and Zwilling 2014), but with a focus on students enrolled in traditional educational institutes of on-site education system; Del Río and Insuasti's (Del Río and Insuasti 2016) analyzed 51 studies published between 2011 and 2016. It concluded that 51.8% of the reviewed literatures used predictive variables from a

combination of higher education indicators obtained from academic performance with other types of attributes. And 37.5% of reviewed literatures used only higher education indicators obtained from academic performance. Del Río and Insuasti (Del Río and Insuasti 2016) inferred that 71.8% from the reviewed papers used the classification method in EDM task. Clustering methods presented 8.9% and association rules presented 7.1% of the studies. In these studies, academic success prediction in degree course and work analysis of student performance was focused on course result not a combination of attributes with a small database. Natek and Zwilling (Natek and Zwilling 2014) reported factors connected to the final results were linked to information on demography and extracurricular activities; which was visible in undergraduate program of bachelor degree in computer science. Major classification algorithms that were used in review studies by Natek and Zwilling (Natek and Zwilling 2014) were: M5P model, RepTree model and J48 model.

Asif (Asif et al. 2017) presented a prominent level of accuracy in academic prediction through using Naïve Bayes and Random Forest Tree algorithms. Attributes used to predict graduation performance of four years undergraduate degree course were pre-university grades and results achieved by students in a course of first and second year of the undergraduate degree course.

Vera L Miguéis (Miguéis et al. 2018) found the best algorithm; Random Forest algorithm, which attained high accuracy prediction based on academic prediction of first year academic performance results. Other algorithms were found too such as decision trees, support vector machines, Naïve Bayes, bagged tree and boosted trees. Also, key factors to predict academic performance in the total undergraduate years in engineering study course were: means of university access and university access examination as well as means of first academic year course.

Vera L Miguéis (Miguéis et al. 2018) proposed early class classification based on student performance results in the first academic year using “multiclass segmentation structure” based on predictive model to achieve academic success.

2.3 Factors that contribute to student attrition

In this section, factors that contribute to student attrition are grouped in three major groups. The first group is factors related to student’s demographics, student’s social factors and student’s psychological factors. The second group is related to student’s prior

performance and academic factors. The third group is related to student's engagement and institutional factors. The last group is financial factors.

2.3.1 Student's demographics, social and psychological factors

University students' dataset includes comprehensive information of students. One type of information is related to students' demographic, which is represented by age, gender, nationality, residence address, marital status. Demographic factors are used in EDM research because it is simple information that is easy to obtain and use to classify by. A research study discovered that there is an association between student demographic factors and student attrition. Delen (Delen 2010) recognized from a public university dataset of 23000 students in USA that age, marital status, ethnicity, gender and permanent address are principal factors in predicting student attrition. On the contrary, another research study found no influence of student demographic on student attrition. Zhang (Zhang et al. 2010) found no influence of demographic factors such as age, nationality and gender on student attrition. The study research was based on undergraduate and postgraduate students' data from Thames Valley University in UK with 4000 student datasets. It shows academic factors were the key aspect in predicting student attrition.

Social factors are not widely spread used in student attrition research, due to the lack of its availability in education dataset. Social factors related to student social attributes such as parent education level and student socio-economic status were found insufficient for prediction model. Fike and Fike (Fike and Fike 2008) discovered that parents' educational level influenced a student's decision in continuing education. Their prediction model of student retention from first and second semester revealed that the father's education influenced a student positively.

Survey studies revealed that psychological factors are considered as major student attrition factors from success. Evans (Evans 2000) concluded that the major psychological factors that lead to students' withdrawal from education were: academic preparedness and study skills, student's commitment to the study and student's intention. The most dominant factor was academic preparedness and study skills which influenced the students to withdraw from a course.

2.3.2 Student's prior performance and academic factors

Student's prior performance is taken into consideration in high school grade because it is the only and most considered education grade used in university enrollment. Del Río and Insuasti (Del Río and Insuasti 2016), Ahmad Tarmizi (Ahmad Tarmizi et al. 2019), Asif (Asif et al. 2017) and Vera L Miguéis (Miguéis et al. 2018) explored the influence of high school grade point average (GPA) in predicting student's withdrawal. Del Río and Insuasti's (Del Río and Insuasti 2016) research shows that GPA is a significant variable used in predicting student's withdrawal. Whereas in Ahmad Tarmizi's (Ahmad Tarmizi et al. 2019) research the GPA does not rank top of the list in withdrawal predictor variables. High school rank was considered as a factor contributing indirectly to prior performance, which related to education performance of the student. Vera L Miguéis (Miguéis et al. 2018) displayed the influence of the high school category attended on student's attrition. This factor exhibits the socio-economic influence and the ability of a high school to deliver good education.

In Australia, ATAR (Australian Tertiary Admission Rank) is one of the ranks used for student's admission to university. There are two contributors to ATAR rank which were tertiary entrance score and tertiary entrance rank. These two contributors effect the quality of education delivered to a student in high school (Miguéis et al. 2018). In the USA, students perform ACT (American College Testing) and SAT (Standardized Admission Test). Delen (Delen 2011) showed the influence of SAT score and how it has a significant effect on student's attrition. Whereas other studies neglected the university entry assessment score. Similarly to SAT Delen (Delen 2011) considered TOEFL assessment's score to evaluate student's English level before university enrollment, but TOEFL was not significant to student's attrition. Adekitan and Noma-Osaghae (Adekitan and Noma-Osaghae 2019) used mathematics and English but discovered that they were not significant. Also, Fike and Fike (Fike and Fike 2008) used completion of mathematics, reading and writing as influences on student's attrition. Enrollment in development courses and successful completion lead negatively to student's attrition. In a comparison between mathematical and reading courses effects, it showed that the completion of reading course successfully influenced the student's attrition negatively.

The student academic factor that influenced student's attrition is: study year in university. Number of research papers showed the importance of current year and its effect

on student's attrition. Zhang (Zhang et al. 2010) found that the majority of students' withdrawals from university were during the first academic year or before the start of the second academic year. Furthermore, the research revealed that the longer a student stayed in study the more likely is the student to continue study course. Although the first year of study has more influence on student's attrition, the second year has an influence as well on a student's withdrawal. Asif (Asif et al. 2017) showed that the second year was linked to student's attrition from course. There were varied factors that lead to student's withdrawal compared to the first year's factors, which were: number of enrolled courses and duration of study course.

Courses choice had a significant influence on student's attrition and was considered as an important characteristic that can be used in prediction. Variables used by Delen (Delen 2010) were major and concentrated, variables used by Zhang (Zhang et al. 2010) was course award, variables used by Adekitan and Noma-Osaghae (Adekitan and Noma-Osaghae 2019) were major and concentrated. Vera L Miguéis (Miguéis et al. 2018) found that many students of engineering field had a higher ratio of course withdrawal, despite the good performance of students in course.

Number of enrolled hours per semester had an important influence on student's withdrawal (Fike and Fike 2008; Delen 2011). Fike and Fike (Fike and Fike 2008) found that students who registered in semester with high number hours were most likely not to withdraw from a course. Delen (Delen 2010) found that the ratio of number of earned hours to registered hours was an important variable in the prediction of a student's withdrawal. Moreover, course attendance variable had an important influence on student's attrition.

Student performance had a direct influence on student's attrition. Delen, Fike and Fike and Zhang (Fike and Fike 2008; Delen 2010; Zhang et al. 2010) showed the significant link of student's marks on student's attrition. Asif (Asif et al. 2017) showed that the link between marks and attrition is negative, students who score lower grades had a higher probability to leave study course. Zhang (Zhang et al. 2010) found that students tend to leave university to another institute when they had a lower entry score and high course marks. Fike and Fike (Fike and Fike 2008) found that a student's probability for attrition increased with the increase of dropped hours per semester.

2.3.3 Student's engagement and institutional factors

With the development of education and introduction of online learning management system LMS, a new variable has been used in student retention data mining studies. This introduced a new source of data measure by researchers to measure a student's engagement in study course. Despite that, few studies used student engagement in student's attrition prediction. Aguiar (Aguiar et al. 2014) used student interaction of login, assignment submission and semester's e-portfolio access to explore student's engagement, which lead to predict course withdrawal. Aguiar (Aguiar et al. 2014) found that student engagement in login and interaction with learning system was associated with student retention. Zhang (Zhang et al. 2010) explored student's usage information of online learning management system and online library system. Zhang (Zhang et al. 2010) found a correlation between usage information and student's withdrawal. Fike and Fike (Fike and Fike 2008) found a negative relation between online courses and the probability of student's withdrawal.

Institutional factors are analogous to the situation with student's engagement in terms of being not widely explored. Tinto (Tinto 1975) found that the integration of institutional factors information in a study model contributed positively to student's attrition prediction. Fike and Fike (Fike and Fike 2008) found that information of university facilities (support services usage) had a negative relation to the probability of student's withdrawal. College variable was used by Delen (Delen 2010), but it had no important influence on student's withdrawal.

2.3.4 Financial factors

Financial variables were used in studying student's attrition patterns. These factors were divided into two groups: the first group is related to education activity and the second group is related to work activity. Delen (Delen 2010) presented a number of variables related to education's financial area which are: scholarship, tuition wavier, getting financial aid, loan and grant. These variables had a noteworthy influence on student's withdrawal. Fike and Fike (Fike and Fike 2008) found a positive relation between financial aid and student's attrition.

Study	Objective	Factors used							Dataset Size – (Training percentage)	Algorithms	Evaluation
		Demographic	Socio-Economic	Prior Performance	Academic performance	Engagement/Interaction	Behavior	Other			
(Adekitan and Noma-Osaghae 2019)	To predict relation between admission criteria and student 1 st year performance	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>				1445 – (70%)	RF, Tree Ensemble, DT, NB, LR, Rprop	Accuracy
(Aguiar et al. 2014)	To study reducing miss-prediction by student engagement	<input type="checkbox"/>			<input type="checkbox"/>				429 – (90%)	NB, DT, LR, HDDT, RF	Accuracy
(Almarabeh 2017)	To predict student performance and find highest accuracy among 5 chosen classifier				<input type="checkbox"/>		<input type="checkbox"/>		225	NB, BayesNet, ID3, J48, MLP	Accuracy
(Alomari, K.M., AlHamad, A.Q. and Salloum 2019)	To predict video game rate categories							<input type="checkbox"/>	2053 – (80%)	GLM, DT, DL, GB, RF, NB	Accuracy, F-Score
(Asif et al. 2017)	To predict student year four achievement and progression			<input type="checkbox"/>	<input type="checkbox"/>				210 -	DT-GI, DT-IG, DT-Acc, RI-IG, 1-NN, NB, NN, RF-GI, RF-IG, RF-Acc	Accuracy

(Beer and Lawson 2017)	Identify factors leads to student attrition							<input type="checkbox"/>	2643	QDM	
(Bucos and Drăgulescu 2018)	To predict student performance from available course data to prove its valuable for future decision							<input type="checkbox"/>	1077	DT CART, ET, RF, LR, C-SVC	TP, TN, Accuracy, F-score
(Delen 2010)	To predict and explain student attrition	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>			<input type="checkbox"/>	16066	ANN, DT, SVM, LR, Ensemble predictors	Accuracy
(Fernandes et al. 2019)	To predict student performance			<input type="checkbox"/>	<input type="checkbox"/>				238,575 (2015) 247,297 (2016)		ROC,
(Fike and Fike 2008)	Analyze predictors for analyzing student retention	<input type="checkbox"/>	<input type="checkbox"/>		<input type="checkbox"/>				9200	BCC, PBCC, Phi-CC	Accuracy
(Miguéis et al. 2018)	To segment student and predict overall student performance								2459	RF, DT, SVM, NB, Ensemble predictors	
(Natek and Zwilling 2014)	To predict student success rate in a course	<input type="checkbox"/>	<input type="checkbox"/>		<input type="checkbox"/>				106 – (70%)	DT (J48, M5P, RT),	Accuracy
(Zhang et al. 2010)	To monitor, analyze and suggest intervention strategies for student prediction	<input type="checkbox"/>		<input type="checkbox"/>	<input type="checkbox"/>			<input type="checkbox"/>	Multiple dataset. Student dataset: 4223 – (66%)	NB, SVM, DT	Accuracy

Table 1: Studied Research Paper Summary

Chapter 3

Methodology

There are several data mining methodologies developed in the knowledge discovery field (Azevedo, 2019). The two most popular methodologies for data discovery are SEMMA (Sample, Explore, Modify, Model, Assess) and CRISP-DM (Cross-Industry Standard Process for Data Mining). The main difference between the two methodologies is that CRISP-DM method has two more steps involving business understanding and deployment.

To maximize the gain from this research on an academic and workplace level, CRISP-DM method has been considered to predict student attrition. Also, CHEDS report template was used in this academic work's knowledge discovery's model to contribute with a DM model that can be implemented on any higher education institute in the UAE.

The design of CRISP-DM method is a general design that can be implemented in a vast range of fields and applications. The CRISP-DM consists of six main phases listed below.

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

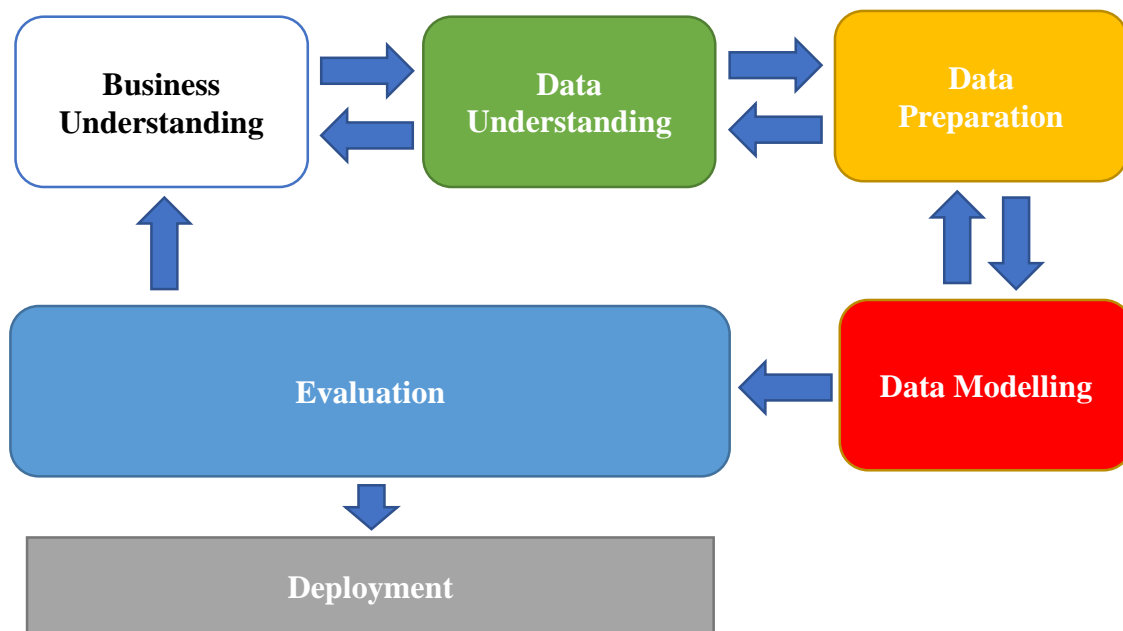


Figure 1: CRISP-DM Workflow

This model does not follow sequential steps or one direction as shown in figure-1; loops and repeating steps are commonly used to successfully complete the knowledge discovery model. The method description of each phase is discussed in the sections below.

3.1 Business Understanding

The aim of this research is to support UAE's higher education sector from student attrition through an academic approach using data mining method. Higher education institutes suffer from students' withdrawal officially or unofficially. Official students' withdrawal is through: the submission of 'study withdrawal request' form or the student not meeting the required performance level. Unofficial students' withdrawal is when the student did not register in the next semester. Student attrition increases the risks on program continuation (may lead to program closure in case the number of students drop less than the threshold of cost breakeven) and effects higher education institutes' reputation negatively. Besides, keeping students who have the possibility of dropping out in a program is simpler and less expensive than attracting new students.

For that, a universal data mining prediction model is the aim in this study to resolve HEI attrition problem. In order to approach that, a unified dataset must be used by all targeted institutes. CHEDS (Central Higher Education Data Store) data repository is a source used in

this research (CHEDS details explained in next phase 3.2.2 data understanding).

The targeted institute in this research is based in the UAE and followed CAA (Commission for Academic Accreditation) certification guideline. It is classified as a college and has only one main premise in Dubai. There are two levels of studies. The first level is Bachelor's degree, where it has two programs: Sharia Bachelor's degree and Law Bachelor's degree. The second level is Master's degree, which is not in the scope of this study.

In every semester, new students' enrollments are around 70 to 100 students out of a total of nearly 500 students' enrollments (mixed of: new, continuing and readmission). Student demography is multi-national and consists of both genders. A high number of students are employees and their study takes place in the evening hours. Study hours are paid by students or partially paid as a scholarship. Admission requires students to pass High school with at least 60% score for Sharia program and at least 80% score for Law program. The Language score required is: 950 in EMSAT, 4.5 in IELTS or 400 in TOFEL.

Dependent variable is student labeled as 'Dropped' with value "Yes" or "No". 'Dropped' in this study is considered when: a student withdrawal is official or unofficial or a student did not register in two consecutive semesters. To approach the aim of the prediction model, the following are a set of research questions:

- 1. Which students are likely to withdraw from the institute?**
- 2. Which variables are more efficient at predicting student withdrawal?**
- 3. How accurate a prediction factor is to rely on to make a decision in an organization?**

3.2 Data Understanding

In this phase, data is collected and explored to understand different segmentations and types of attributes. Attribute types are classified to: First, demographic data that is related to student (Age, Gender, Nationality, resident city); Second, high school data that reflects student's performance in high school and pre-registration exam results (High school score percentage, High school system, High school country, Language score); Third, college performance that presents student performance during higher education (Average registered credit hours in every semester, Cumulated GPA, Student status in registering in every semester) and the last and main attribute is Class Label Withdrawal or "Drop" which is a dependent label.

3.2.1 Data Acquisition

Acquiring students' data and performance data could end up in dealing with multiple databases constructed in different systems. It complicates the data preparation phase. On top of that, with multiple databases, table versions could be changed over the course of time which can cause new attributes to be added to the database or attributes removed or changed. For example, if a language test exam type was changed (from TOFEL to EmSAT), it would complicate mapping old and new attributes to a common attribute because of values range difference.

To avoid the prior complication in this research, it has been decided to select a universal student information data source (CHEDS) where all HEIs follow its guidance when reporting their data to the MOE (Ministry of Education in UAE) using CHEDS portal.

3.2.2 CHEDS Database

The Ministry of Education in the United Arab Emirates owns a centralized data collection of higher education data named: "Central Higher Education Data Store" (CHEDS). All accredited Higher Education Institutes in the UAE have access to the portal and are required to upload unified statistics templates with the institutes' information regularly.

The Ministry's aim from CHEDS is to set up and maintain a data warehouse of all UAE accredited higher education institutes that is a reliable and trusted statistic source of HE in the UAE. CHEDS's data is a foundation for any higher education policy making and development planning on an institute or a federal level.

In the first stage of data exploring, the following five reports were interesting to retrieve data from for the research questions.

1. Student Enrollments Report
2. Student Graduates Report
3. Student Attrition Report
4. Applicants Academic Proficiency Report
5. Applicants Basic Details Report

Academic year	2015-2016		2016-2017		2017-2018		2018-2019		2019-2020		2020-2021	
	Semesters											
Report name	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Student enrollments	Low GPA graduate				High GPA enrolled							
Student graduates											High GPA graduate Low GPA graduate	

Table 2: Comparison of Enrollment and Graduate based on GPA

Due to the small number of students in the participating institute in this research, it has been decided to increase the range of selected academic semesters for data fetching to cover one full study of bachelor's course cycle. Students with high graduate GPA score (Range between 3.7 and 4) completed university study (bachelor's degree) in seven semesters plus summer semesters (which are not included in the dataset). On the other hand, students with lower graduate GPA score (Range between 2.0 and 2.3) completed university study in 10 to 11 semesters. Table–2 exhibits the number of study semesters taken in university, comparing between low GPA graduate students' group and high GPA graduate students' group.

These database files were fetched for 12 semesters to cover one cycle of bachelor's study for all five types of reports, starting from Fall 2015-2016 to Spring 2020-2021. Two types of reports were available for all semesters which are: students' enrollment report and student graduates report. Other reports were not available for all semesters because it was introduced in Fall 2020-2021 academic year. Table – 3 summarizes reports' availability.

Academic year	2015-2016	2016-2017	2017-2018		2018-2019		2019-2020		2020-2021			
Report name	Semester											
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Student enrollments	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student graduates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student attrition	No	No	No	No	No	No	No	No	No	No	Yes	Yes
Applicants - Academic Proficiency	No	No	No	No	No	No	No	No	No	No	Yes	Yes
Applicants - Basic Details	No	No	No	No	No	No	No	No	No	No	Yes	Yes

Table 3: CHEDS Report Availability in Study

The total files received from participating institutes were 24 files (12 files for enrollment and 12 files for graduate). There were two more files for summer semester per academic year, but it was not considered in our study since not all students were enrolled in summer semesters. (It would have affected the machine learning negatively as average credit registered hours would be lowered).

3.3 Data Preparation

This stage is known to researchers as the time-consuming phase. 80% of business researchers' time is consumed on data preparation alone. To obtain high quality data in data mining, four phases in data preparation was required. These phases were implemented on the received database as follows: data cleaning, data integration, data transformation and data reduction explained accordingly.

3.3.1 Data Cleaning

After the data has been collected and analyzed to check available reports from unavailable reports, the next phase started by preparing the data. The first stage of data cleaning is data inspection and cleaning.

The 24 database files were received in Microsoft Excel's format. Files were inspected for information quality or corruption. All files consisted correct information and

the files opened without any problems. From first inspection, it was noticed that column titles were inconsistent in naming across all the files. That was due to the update of data collection templates.

3.3.1.1 Data Comparison

The database collected from CHEDS was compared across academic years. Attributes were reviewed and classified as personnel information and study information. Attributes throughout the dataset were compared and differences were found in attributes’ naming and availability. Manual comparison among database files’ attributes were performed backwards from the latest data file to the oldest data file to figure out attributes’ mapping. Table – 4, summarizes attributes’ changes throughout the targeted academic years.

Comparison between Academic Semesters	Changes Log
Spring (20-21) to Fall (20-21)	No differences in attributes
Fall (20-21) to Spring (19-20)	Complex changes: 1. Disorder and new attributes 2. Discontinuing attributes from older dataset, made tracing attributes from latest and earliest dataset difficult
Spring (19-20) to Fall (19-20)	No differences in attributes
Fall (19-20) to Spring (18-19)	Two new attributes in new data model (<i>fall 19-20</i>)
Spring (18-19) to Fall (18-19)	Attribute title changed with same reference data. 11 new attributes added to newer version and 3 attributes removed from earlier version
Fall (18-19) to Spring (17-18)	Minor differences in older version compare to newer version. 6 attributes were removed in newer version
Spring (17-18) to Fall (17-18)	Minor differences in older version compare to newer version. 4 attributes were removed in newer version
Fall (17-18) to Spring (16-17)	15 attributes added in the newer version. Naming is mostly the same
Spring (16-17) to Fall (16-17)	2 attributes added in newer version
Fall (16-17) to Spring (15-16)	4 attributes added in newer version
Spring (15-16) to Fall (15-16)	No differences in attributes

Table 4: Attributes Changes in Progressinve Years

The number of attributes in CHEDS data collection files are growing, this makes CHEDS dataset a rich database for researchers to consider in Education Data Mining. Table – 5 presents the attribute numbers’ growth.

Academic Year	Attribute per data model
2015-2016 Fall	42
2015-2016 Spring	42
2016-2017 Fall	46
2016-2017 Spring	48
2017-2018 Fall	61
2017-2018 Spring	66
2018-2019 Fall	60
2018-2019 Spring	68
2019-2020 Fall	70
2019-2020 Spring	70
2020-2021 Fall	72
2020-2021 Spring	72

Table 5: Number of CHEDS Attributes in Academic Year

3.3.1.2 Attributes’ Selection and Data Cleaning

Out of 72 attributes in the last available dataset of academic 2020-2021 Spring, 26 attributes were selected to be included in the study. Out of the 26 selected attributes, 8 were found not available in earlier academic period datasets. Table - 6 shows the selected 26 attributes from enrollment dataset and highlights the 9 attributes that were dropped in earlier dataset.

Attributes Name	Missing attributes
Enroll_Academic_Period	
Enroll_Gender	
Enroll_Student_DOB	
Enroll_Health_Fitness_Certificate	From 16-17 Spring and older
Enroll_Marital_Status	From 16-17 Spring and older
Enroll_Nationality	
Enroll_Nationality_of_Mother	From 16-17 Spring and older
Enroll_Home_Emirate	
Enroll_Student_Type	
Enroll_1st_Academic_Period	
Enroll_Student_Level	
Enroll_Student_Degree	From 17-18 Fall and older
Enroll_Mode_of_Study	From 16-17 Spring and older
Enroll_Employment_Status	
Enroll_Required_Academic_Period	From 19-20 Spring and older
Enroll_Required_Credits_Graduation	From 18-19 Fall and older
Enroll_Current_Registered_Credits	
Enroll_Total_Credits_Registered	From 19-20 Spring and older
Enroll_Total_Credits_Cumulated	
Enroll_GPA_Cumulative	
Enroll_Transfer_Institution	
Enroll_Language_Test_Name	
Enroll_Language_Test_Score	
Enroll_High_School_System	
Enroll_High_School_Score	
Enroll_High_School_Country	

Table 6: Selected Attributes in Study

‘Enroll student level’ attribute was used to distinguish master’s students from bachelor’s students, and all master’s students’ data rows were removed from the data as it is not in the study’s scope. Summer semesters’ data were removed from the databases as it affects attributes’ transformation negatively and leads to wrong prediction. Summer semesters’ data were also found in Fall’s database of academic year 2018-2019, 2019-2020 and 2020-2021. CHEDS summer semesters’ codes were (201804, 201904, 201905). Table – 7 summarizes semesters’ codes per CHEDS.

Year code	Description
YYYY-01	Fall semester
YYYY-02	Winter Semester (Not applicable in this college)

YYYY-03	Spring semester
YYYY-04	Summer 1st semester
YYYY-05	Summer 2nd semester

Table 7: Semesters codes

Due to difficulties in attributes' consistency throughout the long academic year data observation, it was decided to select 13 attributes as a final data frame work. It was followed by data integration and then new attributes' creation. Table – 8 lists the final selected attributes and the description of each attribute.

Field Name	Description
Enroll_Gender	The gender of the student
Enroll_Student_DOB	The Date of Birth in YYYY-MM-DD format
Enroll_Nationality	Current country of citizenship of the student as defined by the students' passport
Enroll_Home_Emirate	The Emirate where the student's residence is located as defined on the passport or visa
Enroll_1st_Academic_Period	The first term that the student was registered for his/her current PROGRAM at the institution
Enroll_Student_Level	The students' current level of study
Enroll_Student_Type	Student type as new, continuing, readmission...etc
Enroll_Current_Registered_Credits	Sum of all credits registered for in the current academic period
Enroll_GPA_Cumulative	Cumulative grade point average (CGPA) from the beginning of the student's record until the last enrolled academic period
Enroll_Language_Test_Score	The score attained on the Language proficiency exam
Enroll_High_School_System	The high school SYSTEM for students qualifying from high school system (e.g., UAE, American, British, etc.)
Enroll_High_School_Score	Exit score for high school students
Enroll_High_School_Country	The country from where the applicant obtained his/her last high school diploma/certificate

Table 8: Final Selected Attributes List

3.3.2 Data Integration

Once the final attributes list was defined and selected, data files were required to be integrated into one file for the next modelling phase. To consolidate the 24 files into one file,

Microsoft Excel – Power Query was used. The two most used features to combine the files were: APPEND and MERGE features.

First, a reference list called STUDENT ID was created using APPEND option among all 12 enrollment files. The result was a list of 7250 of student IDs. Duplication option was run to remove duplicated student IDs which caused 5747 student IDs to be removed. The unique student ID list consisted of 1503 students. Out of 1503 students' records, 231 students' records belonged to Master's degree and was removed. Thus, the resulted dataset was 1272 student IDs.

Second, the MERGE option was run using STUDENT ID list as a reference on all 12 enrollment files to create one comprehensive file containing student information in a single entre (single row). Column filter option was used to filter out 60 not required attributes and 12 selected attributes remained. Demography attributes were merged to the list once whereas two attributes were repeated 12 times: attributes related to study performance (*Enroll_Current_Registered_Credits*) and (*Enroll_GPA_Cumulative*).

Third, repeat merge option was used on the 12 graduate files to include graduate GPA value for all graduated students. After this step, the next stage was data transformation.

3.3.3 Data Transformation

The final dataset was constructed. The attribute was required to be transformed to increase prediction accuracy. Likewise, data was transformed to introduce variables for better modeling characteristics results through two transformation stages: initial transformation and automated transformation using RapidMiner Studio Educational version v9.10.011.

3.3.3.1 Initial Transformation

First transformation, was the conversion of 12 values of (*Enroll_Current_Registered_Credits*) to a single value, by finding the average value using MS EXCEL AVERAGE function. At this point, summer semesters' data was omitted as it reduces the average value negatively.

Second, MS EXCEL TEXTJOIN function was used to produce one entry of various demographic attributes. During data fetching (append, merge), a single attribute had multiple repeated values, and with this function it was corrected.

Third, two new attributes were introduced from (*Enroll_Student_Type*) in every semester (12 values found per student), and named 1st Status and 2nd Status. These attributes reflected student status. The first attribute was related to (new, transfer, continuing) status. The second attribute captured any readmission status. Readmission is usually related to students who are unable to continue studying for several reasons and considered as a weak point (Cogan 2011).

Fourth, new feature was created to decide whether a student graduated or “Drop”. The new attribute “Drop” was developed with the following options:

- If student is listed in graduate report, then the value is “No”
- If student is continuing study, then the value is “No”
- If student status for the last two semesters is blank, then the value is “Yes” (to indicate that the student dropped officially or it is considered as student attrition from the institute)

Fifth, student date of birth was changed to age by calculating age based on enrollment semester and date and month were set to 1st of Sep.

Sixth, gender value was changed from Male, Female to 1 and 0 correspondingly.

3.3.3.2 Automated Advance Transformation

Further transformation was needed to increase the quality of dataset before applying data mining algorithm models.

1. Impute missing data

CHEDS has strict procedures to ensure that the database’s high quality is maintained during uploading to the portal. As dataset was retrieved from Ministry of Education’s (CHEDS portal), the percentage of missing information was low as presented in Table-9.

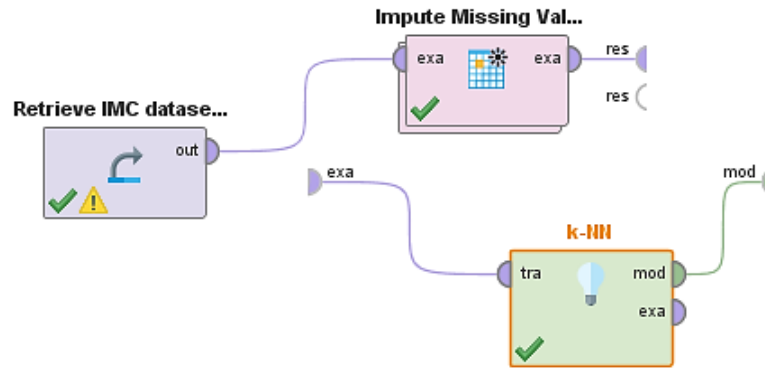


Figure 2: Impute Missing Value Operator

To correct the missing value, ‘Impute Missing Value’ operator was used in RapidMiner. ‘K-Nearest Neighbor’ operator performed the algorithm through sub-process as illustrated in Figure-2 (with k-NN = 5 as default value recommended by the software which showed that most users selected k-NN= 5 for similar implementation).

Attribute	Missing percentage
Gender	0.0%
Student Age	0.0%
Nationality	0.0%
Emirate	0.0%
First_Status	0.0%
Second_Status	0.0%
Language_score	4.1%
HS_Country	0.0%
HS_System	0.8%
HS_Score	0.9%
average_registered_hours	0.2%
Drop	0.0%
GPA	0.1%

Table 9: Missing Information Percentage per Attribute

2. Applied One Hot Encoding

One Hot Encoding was applied for the following attributes (Nationality, Emirates, HS country, HS System, 1st and 2nd status) and decoded to machine state code, which resulted in better performance prediction. Then, old columns were removed and a new file was generated with the code. Figure-3 shows the process used in RapidMiner to apply ‘One-Hot Encoding’ operator. The result of coding the attribute Emirates instead of Dubai was a code= 1000000 and Emirate of Ajman was 0100000, as illustrated in table-10.

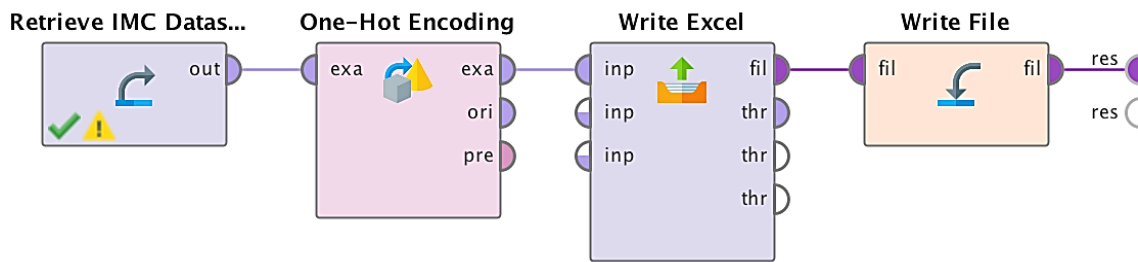


Figure 3: One Hot Encoding Operator

Emirate s		EDuba i	EAjma n	ESharja h	EAlai n	EAbuDha bi	ERA K	EFujaira h
Dubai	=	1	0	0	0	0	0	0
Ajman	=	0	1	0	0	0	0	0

Table 10: One Hot Encoding conversion sample

After Dataset was encoded based on ‘One-Hot Encoding’, the total of attributes reached 68 attributes. City address attribute was converted to 7 attributes. High School attribute was converted to 21 attributes. Nationality attribute was converted to 29 attributes. Table-11, shows new attributes applying advance data preparation before applying data modelling.

Age	HSOman	NJordan
AveRegHr	HSOther	Nkenya
Drop	HSQatar	NMauritania
Gender	HSSaudi Arabia	NMorocco
GPA	HSSudan	NOman
HsScore	HSSyria	NPakistan
EAbuDhabi	HSSyrianBaccalaureate	NPalestine
EAjman	HSTurkiya	NSaudi
EAlain	HSUAE	NSoumali
EDubai	HSYemen	NSri Lanka
EFujairah	NAfghanistan	NSudan
ERAK	NAlgeria	NSyria
ESharjah	NBahrain	NTurkey
HSAfghanistan	NBangladesh	NUAE
HSAlgeria	NCanada	NUnknown
HSAlgerianBaccalureate	NComoros	NUSA
HSAMERICAN	NDominican Republic	NYemen
HSBRITISH	NEgypt	New1

HSEgypt	NGhanaian	Transfer1
HSIndia	NIndia	Continuing1
HSindian	NIran	Continuing2
HSIRAQ	NIraq	LanguageS
HSJordan		
HSMOEUAE		

Table 11: Attributes after Automated Data Processing

3. Data Normalization

Data normalization was needed to scale values to fit in the same values' range. Age, LanguageS, HsScore, AveRegHr and GPA attributes' values were in a different scale compared to attributes that applied with One-Hot Encoding. The method used in normalization was Z-transformation, which subtracted data mean from targeted values then divided values with standard deviation. Figure – 4 shows RapidMiner operator used to perform value scaling Min-Max from 0 to 1.

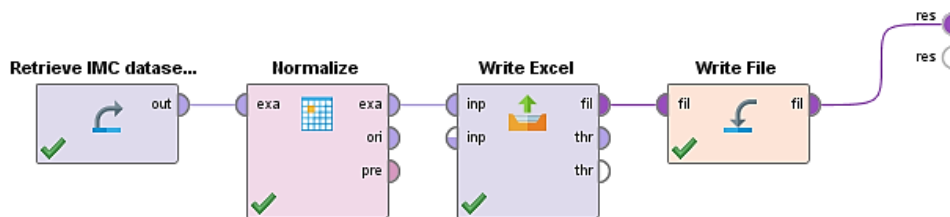


Figure 4: Normalization Operator used

4. Feature selection

Automated feature selection using supervised model was used to find variables which increased the quality and efficiency of the model. Following methods (Info Gain, Information Gain Ratio, Correlation, Chi-Square and Gini index) were used to generate attributes' weight based on class label. Figure – 5 presents the operators used to calculate the attribute weights.

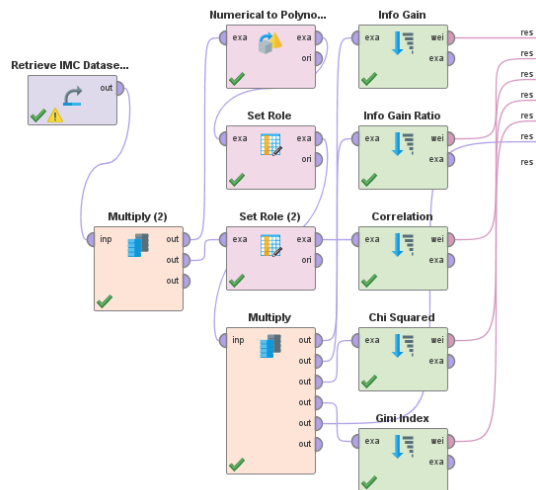


Figure 5: Feature Selection Operators

5. Data overview

The Data Preparation phase was completed after going through Data Cleaning, Data Integration and Data Transformation. Before starting with the Data Modelling phase, Data insight is presented in Figures 6, 7, 8 and 9. Figure – 6 shows the Drop to Graduate/continuing study students are 36% to 64% which is not considered as an unbalanced dataset; because part of the 64% of students (Section 3.3.3.1) were under continuing status (there was a possibility that dropping students were contained in the percentage). Therefore, data balancing was not needed.

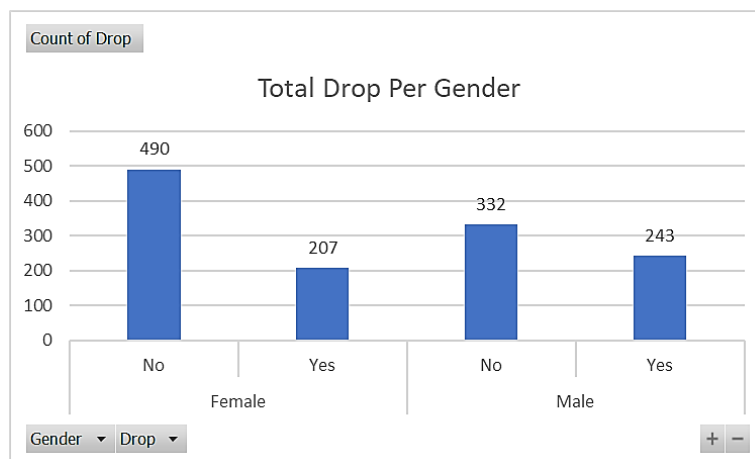


Figure 6: Total Drop per Gender

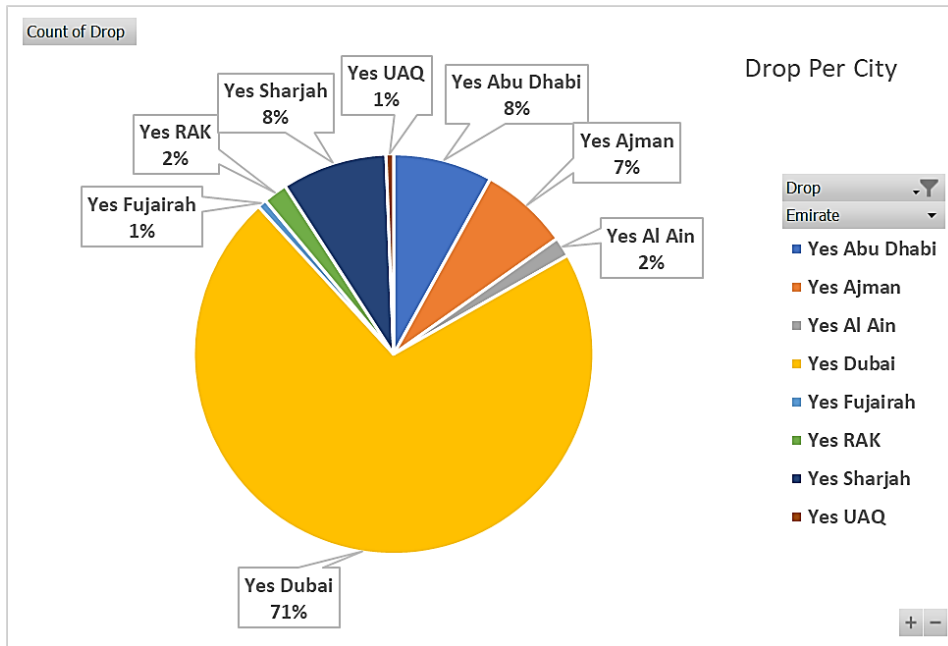


Figure 7: Drop based on City wise

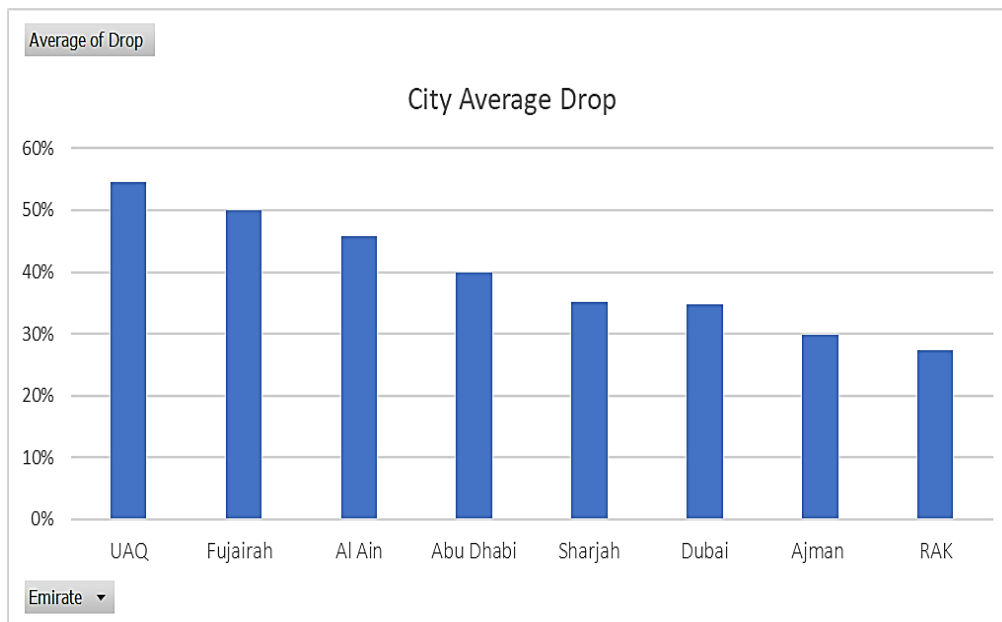


Figure 8: Percentage of Drop per City

Figure – 9 shows percentage of drop per nationality. Four nationalities show 100% drop value because there was only one student’s data from that country. The countries are: Turkiye, Sri Lanka, Ghana and Mauritania.

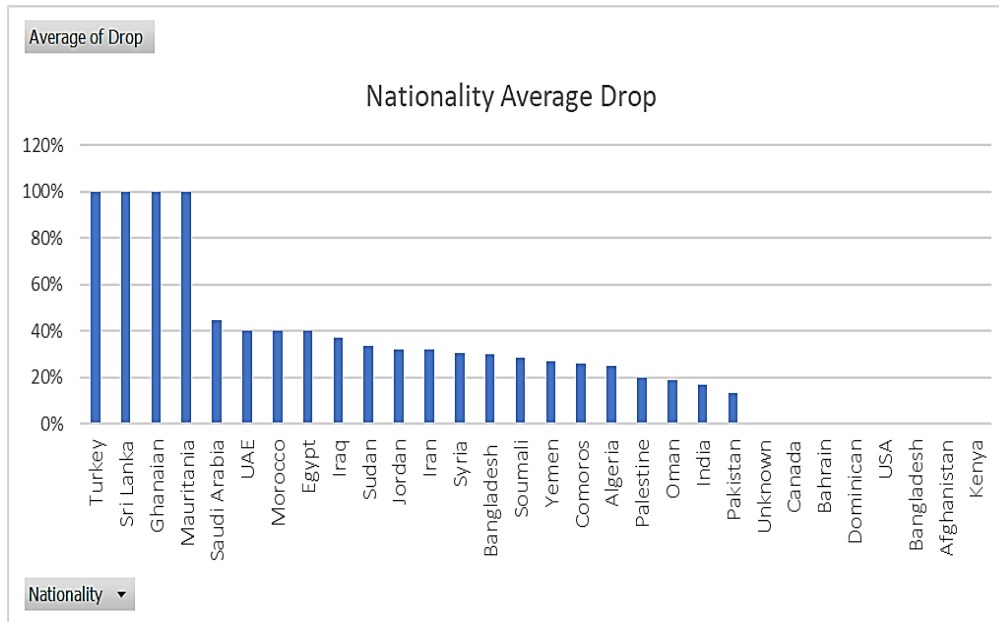


Figure 9: Nationality Drop Percentage

3.4 Data Modeling

This phase was shorter compared to the earlier data preparation phase. It was divided into three sub-phases in our study: first was selecting a modeling technique, second was generating testing design and third was building models (in our case design was made in RapidMiner software).

3.4.1 Modeling Technique

In our study we selected two types of models which are Standard machine learning algorithm and Ensemble machine learning algorithm.

3.4.1.1 Standard machine learning algorithm

1. Decision Tree

It is an algorithm that gives a decision on values based on class. It is presented as a tree with branches of nodes that produces a decision. Each node acts as a splitting node based on data attributes. This rule is used for classification to sperate values based on class. Creation and splitting of nodes is continued until the criteria is met.

Prediction is figured out of class label based on majority of leaf; and numerical

estimated value is based on average value of leaf. Label class attribute is required for RapidMiner to process it in nominal for classification and numerical in regression.

Optional parameter (Criterion) was used and its results were measured. Criterion parameter had different optimization options which controlled value splitting based on:

- i. **Information Gain:** It is a method where attributes with enormous number of values are selected; and it performs splitting on the least entropy.
- ii. **Gain Ratio:** It is a method that reduces the bias of multi value attributes and takes into consideration the ratio of information gain and count of branches when selecting an attribute.
- iii. **Gini Index:** It measures the impurity of nodes based on probability.
- iv. **Accuracy:** It is a method that increases the accuracy of result tree based on selection of attribute splitting.

2. Naïve Bayes

The main assumption value of class label and value of other attributes are independent. The operator in RapidMiner uses Gaussian probability to generate attribute data model.

3. k- Nearest Neighbor

k-NN is an algorithm comparing unknown data to k training data, which is the nearest neighbor to targeted input. The first stage is to find the nearest k of training data by calculating the distance in n -dimensional. In our case, this calculation was automated using RapidMiner while setting the input value to $k=10$ because it produced the least amount of error percentage. The prior will be presented in the Result and Discussion Chapter. The second stage is to classify unknown data based on majority vote of nearest neighbor.

4. Logistic Regression

It is a statistical model taking probabilities of events by having log-odd in a linear combination with independent variable (Delen 2011). In case of logistic regression of binary, a single binary dependent variable is coded by variable indicator with two label values 0 and 1.

5. Support Vector Machine

It is an automated algorithm that assigns label to an object based on an example (Noble 2006). SVM algorithm can predict students' withdrawal by processing and applying model on large amount of data of dropout students and continuing students.

6. Deep Learning

It is a multi-layer of feed forward processing units that consists of artificial neural network and creates a trained model (Hernández-Blanco et al. 2019). The model uses back feed (backpropagation) to calculate errors to adjust weight of each connection in the neural network and improve the results.

3.4.1.1 Ensemble Machine Learning Algorithm

Ensemble learning technique is an improved approach for classification accuracy. This technique is performed by constructed aggregated based classifier. The following ensemble were used in RapidMiner and their results are compared in the Result and Discussion Chapter.

1. Random Forest

It is a model constructed on the bagging method of using different bootstrap sample of training data to learn decision tree. Random Forest improves generalization results by building decorrelated decision tree. The difference between Random Forest and Bagging is that RF selects nodes in every splitting tree based on a small set of random attributes.

2. Stacking

From its name, it is an approach that combines multiple different models and divides the dataset into multiple subsets; the training data uses stratified sampling (Sikora, Riyaz). Stacking procedure comes in three steps: first, dividing training data into two separate sets; second, training multiple classifiers on 1st set and test on 2nd set; third, using prediction as input and correct response as output.

3. Voting

The voting model uses multiple classifiers to generate predictions based on average

predictions from the multiple joint classifiers. In our study four classifiers were used which are: Decision Tree, Naïve Bayes, k-NN and Neural Network

4. Bagging

Bagging is also known as bootstrap aggregation. This technique uses continuous sampling with replacement from the dataset based on probability distribution uniform. Bagging method is favorable if the base classifier is unstable for error reduction.

5. Boosting

Boosting technique assigns weight to training example and may change at end of boosting round weight of each example based on performance. Boosting focuses on hard to classify examples.

6. Gradient Boosted Tree

GBT is an ensemble classification tree model. It performs forward learning technique that results in a gradually improving predictive estimation. In our model a series of decision tree (default value=50 recommended by RapidMiner) were applied with weak classification algorithm that gradually changed data. On the other hand, boosting trees were applied to increase the accuracy.

3.4.2 Testing Design

In the modeling phase of our research, after selecting the modeling technique came the second sub-phase. It included designing the testing procedure to ensure the model quality and validity.

The testing phase requires a training data partition and testing data partition. With the help of data mining automation software, split validation operator was selected. The split ratio was 70% of dataset training and 30% was testing. Sampling type was set to automatic. The mode was stratified as default, if it was not applicable, the shuffled would be used.

3.4.3 Building Model

The third sub-phase in data modeling was building data model algorithm for classification prediction based on the testing design. Each considered classification model in section 3.4.1 has setup parameters which are illustrated in following tables.

3.4.3.1 Decision Tree

Model	Decision Tree
Parameters	
Criterion	Gain ratio/Info Gain/Gini Index/Accuracy
Maximal depth	10
Apply pruning	Yes
Confidence	0.1
Apply pre-pruning	Yes
Minimal gain	0.01
Minimal leaf size	2
Minimal size for split	4
No. of pre pruning alternatives	3

Table 12: DT Set Parameters

3.4.3.2 k – Nearest Neighbor

Model	k nearest neighbour
Parameters	
k	10
Measure Type	Mixed Measures
Mixed Measures	Mixed Euclidean Distance

Table 13: k -NN Set Parameters

3.4.3.3 Support Vector Machine

Model	Support Vector Machine
Parameters	
kernel type	dot
C	0
Convergence epsilon	0.001

Table 14 SVM Set Parameters

3.4.3.4 Deep Learning

Model	Deep Learning
Parameters	
Activation	Rectifier
Epochs	10
Compu variable importance	Yes

Table 15: Deep Learning Set Parameters

3.4.3.5 Random Forest

Model	Random Forest
Parameters	
Number of Trees	100
Criterion	Gain ratio
Maximal depth	10
Guess subset ration	Yes
Voting strategy	Confidence Vote

Table 16: RF Set Parameters

3.4.3.6 Bagging

Model	Bagging (DT)
Sample Ratio	0.9
Iterations	10
Parameters	
Criterion	Gain ratio
Maximal depth	10
Apply pruning	Yes
Confidence	0.1
Apply pre pruning	Yes
Minimal gain	0.01
Minimal leaf size	2

Table 17: Bagging Set parameters

3.4.3.7 Boosting

Model	Boosting (DT)
Sample Ratio	1.0
Iterations	10
Parameters	
Criterion	Gain ratio
Maximal depth	10
Apply pruning	Yes
Confidence	0.1
Apply pre pruning	Yes
Minimal gain	0.01
Minimal leaf size	2

Table 18: Boosting Set Parameters

3.4.3.8 Gradient Boosting Tree

Model	Gradient Boosted Tree
Split	Relative
Split Ratio	0.7

Sampling Type	Automatic
Parameters	
Number of Tree	50
Maximal depth	5
Min rows	10.0
Min Split improvement	1.0E-5
Number of bins	20
Learning rate	0.01
Sampling Rate	1.0
Distribution	AUTO

Table 19: GBT Set Parameters

3.5 Model Assessment and Evaluation

After the completion of data modeling, assessment and evaluation was needed. Its purpose was to assess the planned model based on its potential performance as in design testing plan, also to produce the targeted prediction.

In the aim of evaluating classification model results, seven assessment measures were selected, and each measure had its own characteristic in evaluating the result (Francis and Babu 2019).

Accuracy is a measure that observes error and measures how close the result measurement to the correct value (Accuracy 1994). Accuracy is measured through the equation below:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{False Positive} + \text{True Negative}}$$

Precision is a measure that works similarly to Accuracy by observing the error, but precision measures how close the result is to another result (Accuracy 1994). Precision is measured through the equation below:

$$\text{Precision} = \frac{\text{True Postive}}{\text{True Positive} + \text{False Positive}}$$

Accuracy and precision are assessing the error in the model's result. Figure-10, presents the difference between these two-measurement metrics.

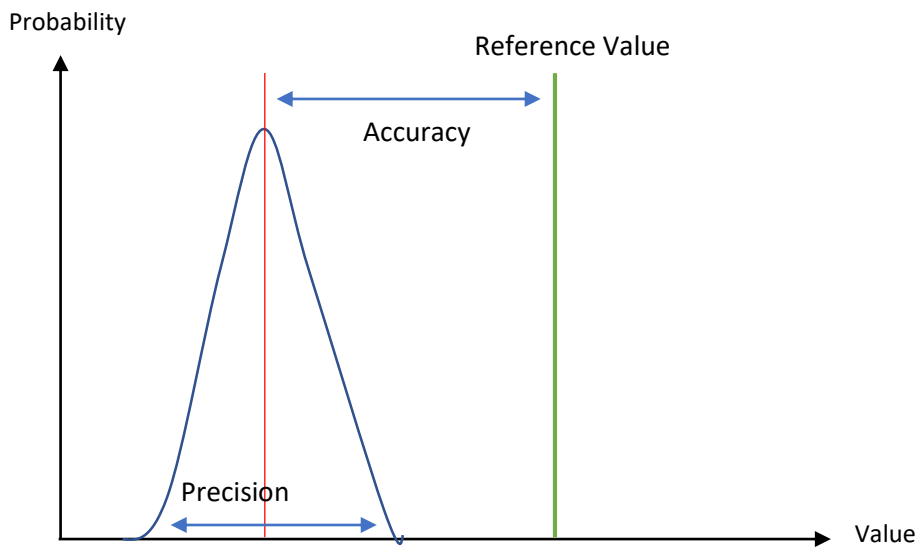


Figure 10: Accuracy and Precision Difference

Recall, is a measure of model ability that finds positive samples, also known as True Positive Rate (TPR) or Sensitivity:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

TNR, True Negative Rate or Specificity is a measure that assesses the accuracy of true negative through the equation below:

$$\text{TNR} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

F-Measure, is a measurement of test accuracy. It is measured by precision and recall value (Powers, David M W) as illustrated in the equation below:

$$\text{F - Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under the Curve (AUC), is a measure of classification performance in threshold level. AUC value varies between 0 and 1, where prediction is correct when AUC has higher values. Figure-11 illustrates the main four shapes of Receiver Operator Characteristic (ROC) based on Area Under Curve value. AUC=0.7 is considered as a good

result for the classification.

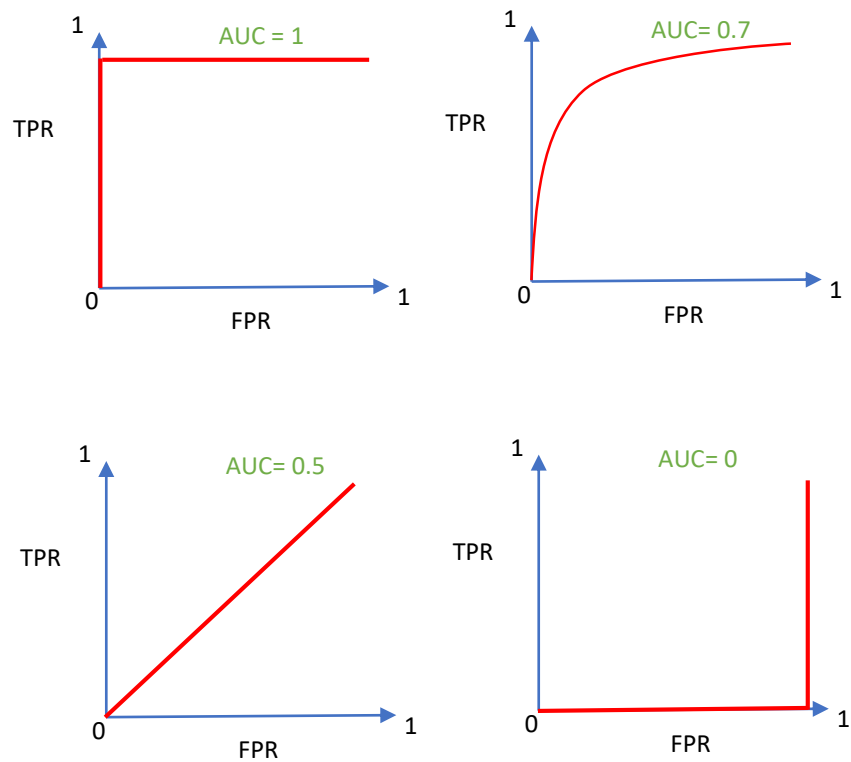


Figure 11: ROC variant based on AUC Value

3.6 Data Deployment

The classifier prediction model was developed to fulfill research question. Deployment steps followed CRISP-DM framework. The model aimed to predict students' withdrawal in governmental higher education institutes in the UAE using CHEDS database that belonged to the Ministry of Education in the UAE. The software used in data modeling was RapidMiner. Data split for training and testing was 70:30.

Chapter 4

Result and Discussion

After successfully proceeding through CRISP-Data Mining workflow, and completing the most complex phases of data preparation and data modeling, it was time to evaluate the model classification output in reference to study motivation research's questions.

The goals of this study were to: identify students that are likely to withdraw from institute, identify the most efficient variable at students' withdrawal prediction and estimate the accuracy of prediction.

4.1 Research Question 1 and 2

1. Which students are likely to withdraw from the institute?
2. How accurate a prediction factor is in making a decision in an organization?

To answer these questions, a classification model was required to be built to predict the class label student attrition. They were two classes from the prediction model: yes (withdraw) or No (Continuing/Graduate).

4.1.1 Best Performance Model

The best prediction classification from standard classifiers with an Accuracy percentage that reached **85.60%** was Decision Tree Model with sub configuration split criterion: Accuracy. This model showed that **94.44%** of predictions were correct from positive class of prediction student withdrawal (Precision), which was a good sign that demonstrated the reliability of the prediction's results; with a recall score of **62.96%**. It was important to find how both precision and recall were harmonious. For that, F1-measure was calculated and the result was **75.56%**. The model performed excellently in finding the

negative class (Continuing/Graduate) based on truly negative, with a result of **97.98%** in Specificity. As for finding the positive class (Withdrawal) based on truly positive, the model got good results with a result of **62.96%** in Sensitivity. Table-21, shows model performance.

Result of Accuracy: accuracy: 85.60%			
	true No	true Yes	class precision
pred. No	242	50	82.88%
pred. Yes	5	85	94.44%
class recall	97.98%	62.96%	
Result of Precision: precision: 94.44% (positive class: Yes)			
	true No	true Yes	class precision
pred. No	242	50	82.88%
pred. Yes	5	85	94.44%
class recall	97.98%	62.96%	
Result of Recall: recall: 60.00% (positive class: Yes)			
	true No	true Yes	class precision
pred. No	242	50	82.88%
pred. Yes	5	85	94.44%
class recall	97.98%	62.96%	

Table 20: DT Model Performance

The Area Under the Curve of Receiver Operator Characteristic scored **0.805**, which was a good sign in finding True Positive from True Negative class. Figure-12 presents

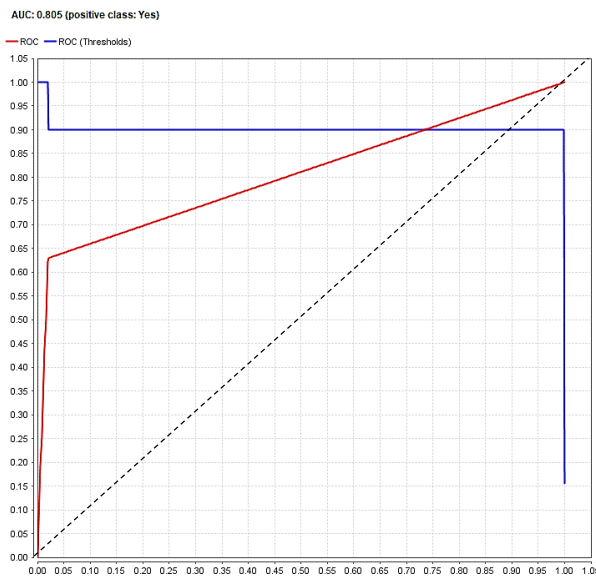


Figure 12:Decision Tree AUC

decision tree ROC curve. The larger the area under the red curve is, the better the performance. It is judged by the area between the red curve and the dotted diagonal line. The value of the diagonal line is AUC=0.5.

Discretization was applied to the decision tree. The improvement performance effect on accuracy was a fraction. The improvement performance effect on Recall was noticeable.

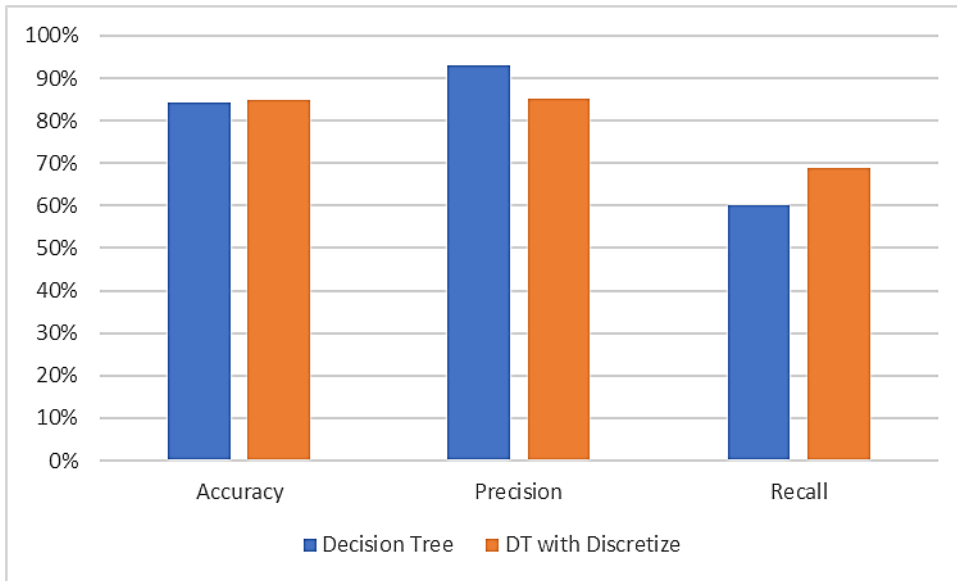


Figure 13: Illustration of DT and with Discretize

But precision performance was lowered due to the nature of discretization in loss of information. Table 22 presents the details in performance changes. And figure 13 illustrates the difference in precision.

Split Criterion: Gain Ratio	Decision Tree	Decision Tree with Discretize Binning
Accuracy	84.29%	84.82%
Precision	93.10%	85.32%
Recall	60.00%	68.89%
AUC - Decision Tree with Discretize Binning		

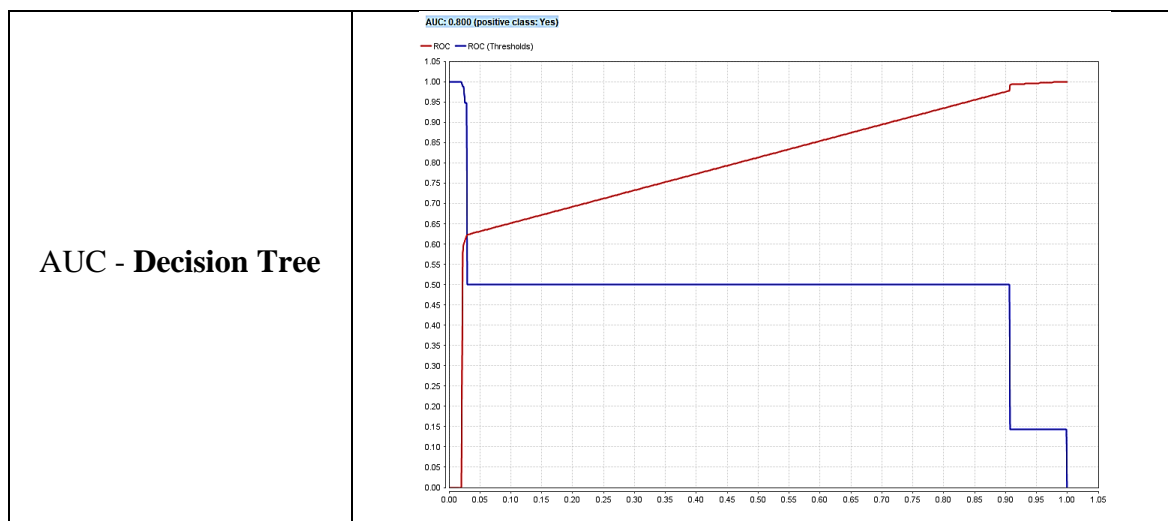


Table 21: Comparison between DT and with Discretize

Multiple split criteria were applied on the decision tree model without the discretized. The best result in accuracy, precision, F-measure and TNR found in split criteria was Accuracy. Whereas Info Gain performed better in Recall and AUC. Table – 23 presents the model performance with split criteria (Gain ratio, Information gain, Gini index and Accuracy).

Decision Tree							
Split criterion	Accuracy	Precision	Recall	F-Measure	AUC	TPR	TNR
Gain Ratio	84.03%	90.22%	61.48%	72.97%	0.800	60.00%	97.57%
Information Gain	84.55%	87.25%	65.93%	75.11%	0.815	65.93%	94.74%
Gini Index	82.46%	84.00%	62.22%	71.49%	0.742	62.22%	93.52%
Accuracy	85.60%	94.44%	62.96%	75.56%	0.805	62.96%	97.98%

Table 22: DT performance of Split Criteria

On the other hand, the best prediction classification from Ensemble classifiers was **Random Forest model which is considered the best model**. The area under the curve scores the highest in all model classifiers AUC=0.878. As presented in table – 24 Random Forest performance details.

Model	Accuracy	F-Measure	AUC	TPR	TNR
Random Forest	84.82%	72.64%	0.878	57.04%	100.00%

Table 23: Best AUC Random Forest Performance

4.1.2 Least Performance Model

The least performing model in prediction classification was the Naïve Bayes model. Accuracy scored **40.05%** and precision scored **36.65%**. These results show that the model did not meet the satisfactory requirement of prediction, where more than 60% of predictions for students' withdrawal were false.

k-Nearest neighbor model performance was low. The least error percentage was found when k-value was equal to 10 (as table – 24 illustrates) compared to other k values. The performance of KNN model was low, almost to 0.5 the diagonal line (AUC=0.6) which led to an unsatisfactory model. Table 25 shows the comparison of area under the curve between Naïve Bayes and k nearest neighbor.

KNN Value	1	5	10	15	20	30
Error %	37.17%	35.60%	31.68%	33.25%	32.98%	32.46%

Table 24: k Value and Error Percentage

Model	AUC
--------------	------------

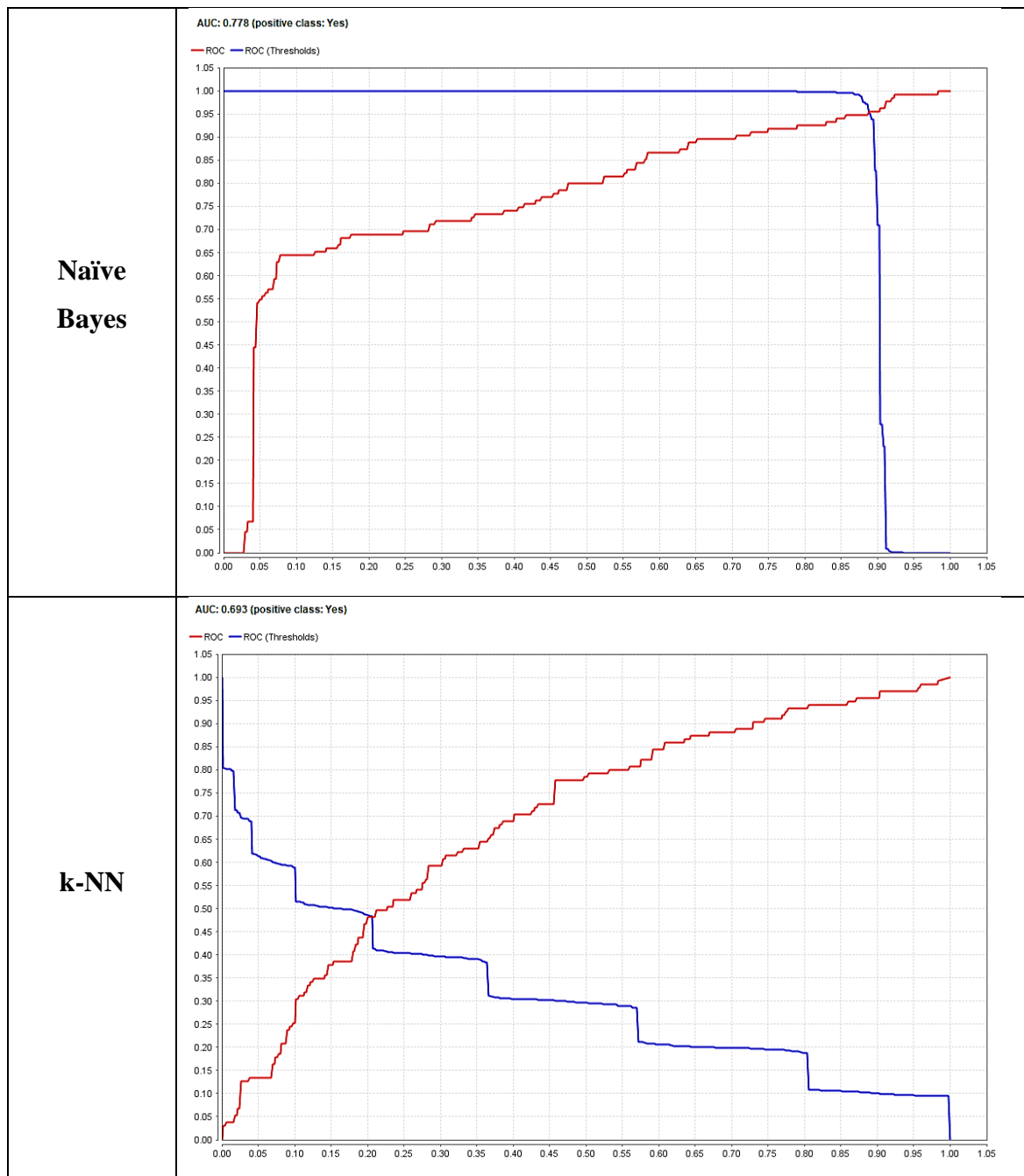


Table 25: Comparison of AUC NB and KNN

4.1.3 Summary of Planned Models

Table – 27 presents performances of standard classifiers and ensemble classifiers. The top area under the curve was scored by Random Forest model AUC=0.878. Naïve Bayes and k nearest neighbor performed the lowest and they are not recommended to be considered. The remaining classifiers performed particularly good to excellent.

Type	Model	Accuracy	F-Measure	AUC	TPR	TNR
Standard Classifiers	Decision Tree	85.60%	75.56%	0.805	62.96%	97.98%
	Naïve Bayes	40.05%	52.98%	0.778	95.56%	9.72%
	KNN	68.32%	46.22%	0.693	38.52%	84.62%
	Logistic Regression	82.20%	71.43%	0.798	62.96%	92.71%
	SVM	83.25%	73.11%	0.817	64.44%	93.52%
	Deep Learning	81.94%	72.29%	0.831	66.67%	90.28%
Ensemble Classifiers	Random Forest	84.82%	72.64%	0.878	57.04%	100.00%
	Stacking	83.77%	73.73%	0.810	64.44%	94.33%
	Bagging	84.55%	73.54%	0.831	60.74%	97.57%
	Voting	84.29%	72.22%	0.784	57.78%	98.79%
	Boosting	84.82%	74.11%	0.802	61.48%	97.57%
	Gradient Boosted Tree	84.55%	74.89%	0.861	65.19%	95.14%

Table 26: Summary All Classifiers' models

Figure – 14 presents selective classifiers to show the performance difference. The selected from top to down are: Random Forest, Deep Learning, Logistic Regression, Naïve Bayes and k Nearest Neighbor.

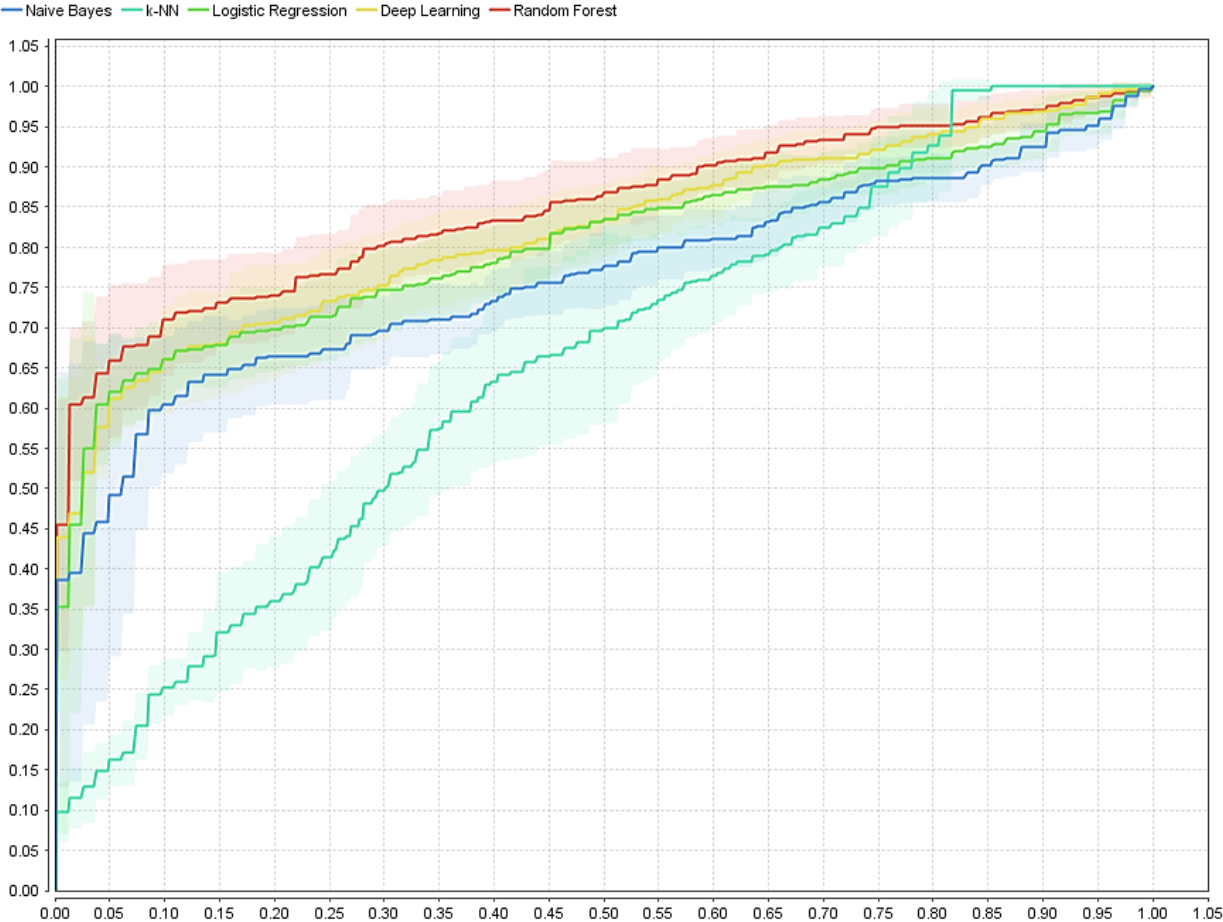


Figure 14: Summary of AUC Diagram

4.2 Research Question 3

3. Which variables are more efficient at predicting a student’s withdrawal?

To answer this question, feature weight analysis was required to be performed to measure the weight for each independent variable compare to class label. There are five metrics to calculate the weight: Correlation, Gini Index, Gain Ratio, Information Gain and Chi-Square.

From the 67 attributes, the top five attributes were: GPA, AveRegHr, Continuing1, New1 and Gender presented in figure – 15. GPA was the variable that had the most effect on the prediction model. The weight of variables was similar in all metrics in terms of ranking. The values of weight in figure – 15 was converted to percentage for the comparison.

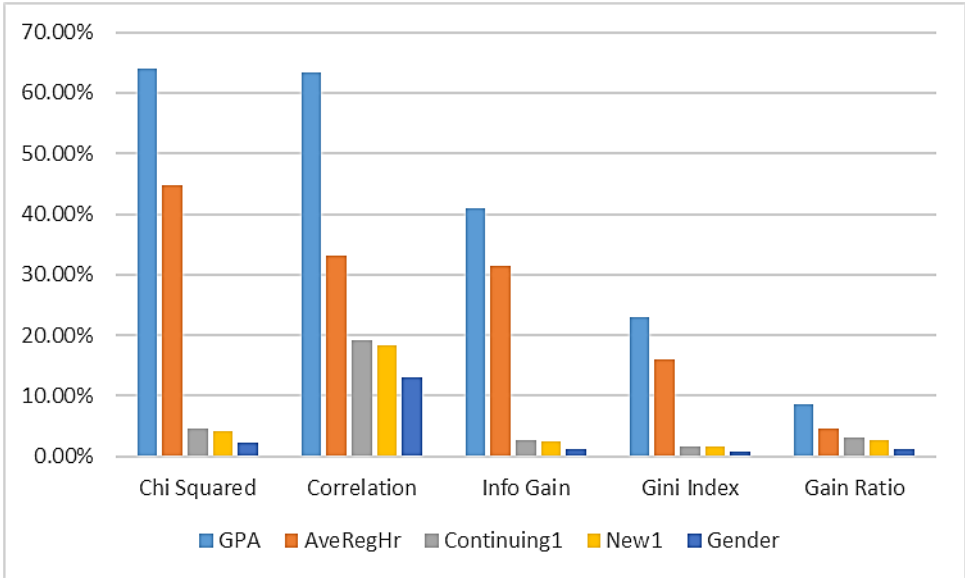


Figure 15: Top Five feature Weight Attributes

Table 27 illustrates the relation among high weight features. Where GPA was high correlated with Average Register Credit Hours. Also, the status of student’s study “continuing” contributed in the GPA. Figure – 16 illustrates the correlation matrix of all attributes.

	GPA	AveRegHr	Continuing1	New1	Gender
GPA	1.000	0.426	0.241	-0.242	-0.162

AveRegHr	0.426	1.000	0.147	-0.161	0.038
Continuing1	0.241	0.147	1.000	-0.929	-0.054
New1	-0.242	-0.161	-0.929	1.000	0.004
Gender	-0.162	0.038	-0.054	0.004	1.000

Table 27: Top Variable Weight Correlation

Table – 28 presents the feature weight of attributes based on each calculation. Only 34 of attributes are listed in the table. Figure – 17 is a Decision Tree model that shows the top predictors were GPA and AveRegHr.

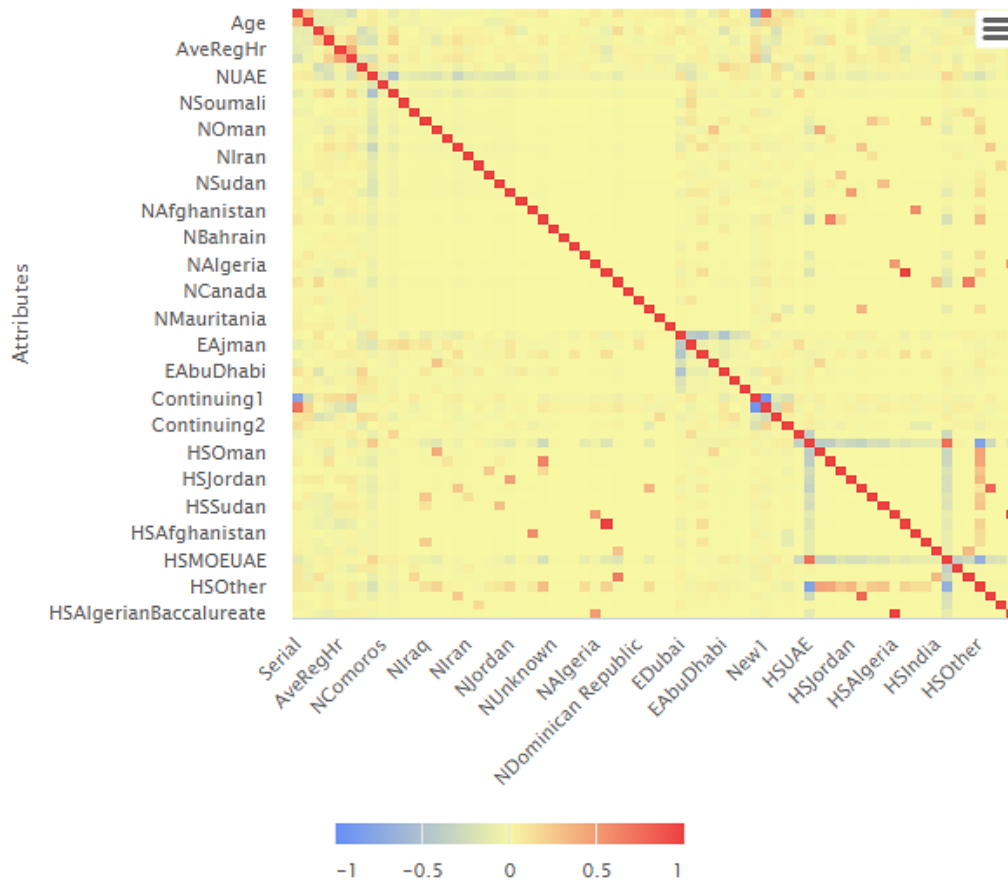


Figure 16: Attribute Correlation Matrix

Sr	Correlation		Gini Index		Gain Ratio		Info Gain		Chi Squared	
	Attribute	Weight	Attribute	Weight	Attribute	Weight	Attribute	Weight	Attribute	Weight
1	GPA	0.6329	GPA	0.2297	NTurkey	0.1396	Age	0.4253	GPA	638.9252
2	AveRegHr	0.3315	Age	0.2128	HSTurkiya	0.1396	GPA	0.4087	Age	592.0964
3	Continuing 1	0.1911	AveRegHr	0.1610	NGhanaian	0.1276	AveRegHr	0.3148	AveRegHr	447.9366
4	New1	0.1828	HsScore	0.1537	NMauritania	0.1276	HsScore	0.2881	HsScore	427.5970
5	Gender	0.1308	LanguageS	0.0614	NSri Lanka	0.1276	LanguageS	0.1163	LanguageS	170.8446
6	NUAE	0.1223	Continuing 1	0.0167	HSIndia	0.1276	Continuing 1	0.0275	Continuing 1	46.4602
7	HsScore	0.0876	New1	0.0153	GPA	0.0868	New1	0.0249	New1	42.5123
8	Age	0.0767	Gender	0.0078	HSJordan	0.0689	Gender	0.0123	Gender	21.7486
9	NYemen	0.0720	NUAE	0.0068	NAfghan	0.0620	NUAE	0.0109	NUAE	19.0225
10	NOman	0.0559	NYemen	0.0024	HSIRAQ	0.0620	NYemen	0.0039	NYemen	6.6028
11	NTurkey	0.0536	NOman	0.0014	NUnknown	0.0586	HSJordan	0.0030	NOman	3.9696
12	HSTurkiya	0.0536	NTurkey	0.0013	NBahrain	0.0536	NOman	0.0025	NTurkey	3.6591
13	HSJordan	0.0509	HSTurkiya	0.0013	Nkenya	0.0536	NTurkey	0.0024	HSTurkiya	3.6591
14	NPakistan	0.0504	HSJordan	0.0012	NCanada	0.0536	HSTurkiya	0.0024	HSJordan	3.3002
15	HSOther	0.0423	NPakistan	0.0012	NDom. Rep	0.0536	NPakistan	0.0021	NPakistan	3.2264

16	HSAMERIC	0.0410	HSOther	0.0008	NUSA	0.0536	NAfghan	0.0015	HSOther	2.2718
17	NIndia	0.0382	HSAMERIC	0.0008	HSAgeria	0.0536	HSIRAQ	0.0015	HSAMERIC	2.1407
18	NGhanaian	0.0379	NIndia	0.0007	HSAfghan	0.0536	HSOther	0.0014	NIndia	1.8551
19	NMauritania	0.0379	NGhanaian	0.0007	HSQatar	0.0536	NIndia	0.0012	NGhanaian	1.8281
20	NSri Lanka	0.0379	NMauritania	0.0007	HSAIlg.Bac	0.0536	NGhanaian	0.0012	NMauritania	1.8281
21	HSIndia	0.0379	NSri Lanka	0.0007	Age	0.0504	NMauritania	0.0012	NSri Lanka	1.8281
22	NAfghan	0.0360	HSIndia	0.0007	AveRegHr	0.0462	NSri Lanka	0.0012	HSIndia	1.8281
23	HSIRAQ	0.0360	NAfghanistan	0.0006	HsScore	0.0384	HSIndia	0.0012	NAfghanistan	1.6462
24	HSOman	0.0336	HSIRAQ	0.0006	Continuing I	0.0302	HSAMERIC	0.0012	HSIRAQ	1.6462
25	HSSyr.Bac	0.0324	HSOman	0.0005	New I	0.0265	NUnknown	0.0010	HSOman	1.4351
26	EAjman	0.0312	HSSyr.Bac	0.0005	NPakistan	0.0231	HSOman	0.0009	HSSyr.Bac	1.3312
27	EAlain	0.0303	EAjman	0.0004	LanguageS	0.0205	EAjman	0.0007	EAjman	1.2405
28	EFujairah	0.0298	EAlain	0.0004	HSSyr.Bac	0.0192	HSSyr.Bac	0.0007	EAlain	1.1698
29	EAbuDhabi	0.0297	EFujairah	0.0004	NIndia	0.0154	NPalestine	0.0006	EFujairah	1.1330
30	NUnknown	0.0294	EAbuDhabi	0.0004	NOman	0.0147	EAlain	0.0006	EAbuDhabi	1.1255
31	NPalestine	0.0286	NUnknown	0.0004	HSOman	0.0126	EAbuDhabi	0.0006	NUnknown	1.0966
32	NSyria	0.0269	NPalestine	0.0004	HSAMERIC	0.0125	EFujairah	0.0006	NPalestine	1.0425

33	NComoros	0.0264	NSyria	0.0003	Gender	0.0124	NSyria	0.0005	NSyria	0.9202
34	HSUAE	0.0249	NComoros	0.0003	NUAE	0.0112	NComoros	0.0005	NComoros	0.8843

Table 28: Independent Variable Weight to Class Label

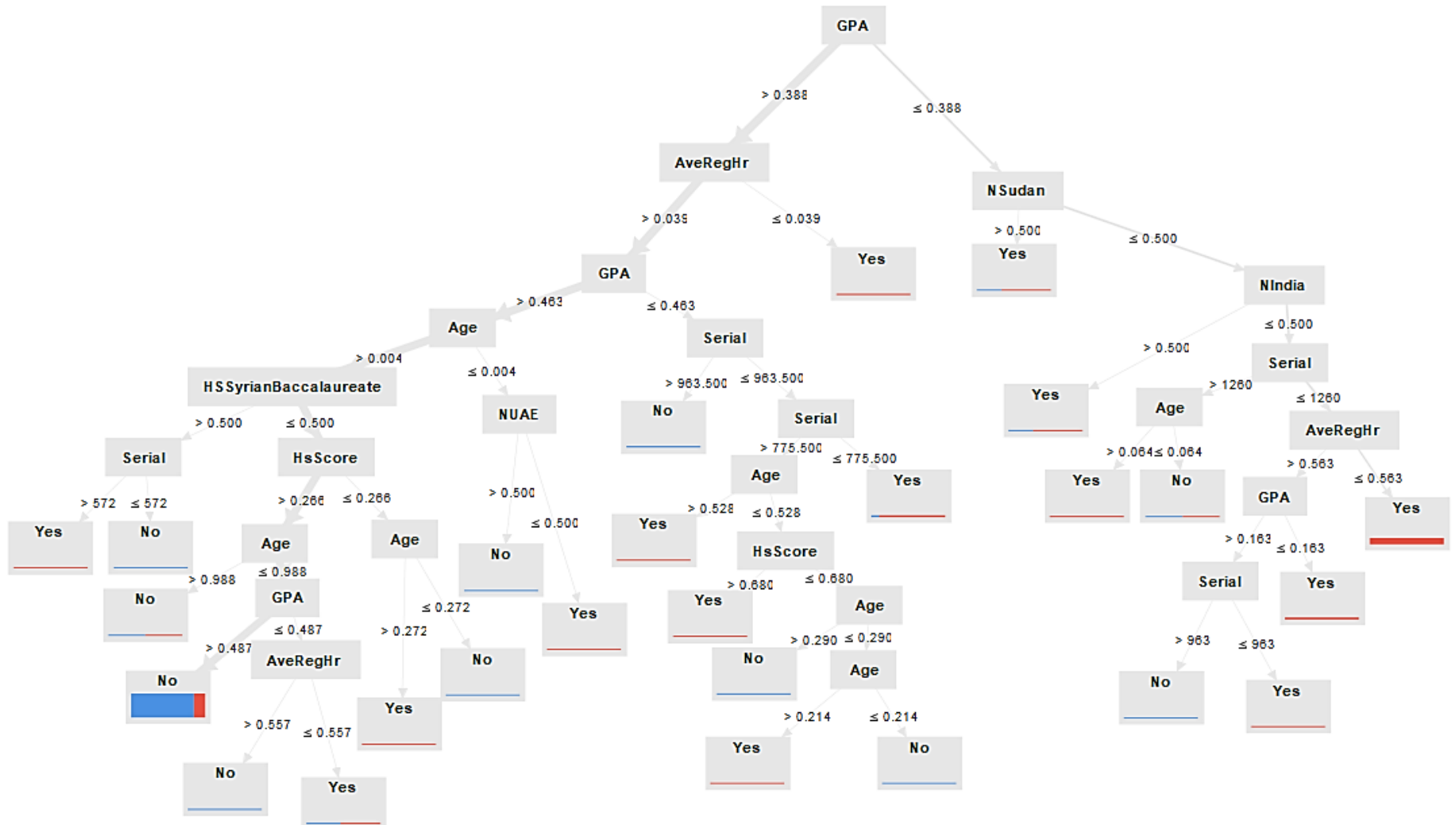


Figure 17: Decision Tree Model

Chapter 5

Conclusion

Student attrition from study course affects a country's development level, and consumes resources and time. It is known as dropout or withdrawal when a student does not complete the planned study course and is not awarded with program certification. Withdrawal is considered official when performance level is not met for the next semester or the student submits a request to withdraw from a program. In this research, withdrawal was considered unofficial when a student did not show in two consecutive semesters.

As far as my knowledge goes and based on Google's scholar search results, this research study provides a unique insight that is not found in earlier research studies that were done in the UAE. The input database was fetched from CHEDS statistical portal that belongs to the Ministry of Education in UAE. This research adds a valuable contribution to the EDM field in the UAE, as this classification prediction model is applicable on other Higher Education Institute in the UAE (HEIs that follow accreditation of CAA).

The fetched database belongs to a Dubai's governmental higher education institute. The dataset consists of 1272 students' data. Student demography was multinationals from both genders.

Finding and Contribution

The aim of the study was to review research papers and build a prediction classification model to: identify which students were more likely to withdraw from the institute, identify which variables were more efficient at predicting a student's withdrawal, and estimate a reliable prediction accuracy to make a decision in an organization.

CRISP-DM method was followed in this research to develop prediction algorithms. 12 algorithms were applied on the dataset. Six were from standard machine learning algorithms and another six were from Ensemble machine learning algorithms. Area under the Curve was used plus other metrics to assess the model's performance.

Result and discussion chapter illustrates the output performance of the models and

concludes our research study by summarizing the answers of the following research questions:

- 1. Which students are likely to withdraw from the institute?**
- 2. How accurate a prediction factor is in making a decision in an organization?**

A student who had low GPA, average register credit hours and fluctuating student enrollment's status was more likely to withdraw from course study. From the 12-prediction classification models, Random Forest from ensemble classifiers illustrated the highest performance in prediction, and scored 87.8% in AUC with an accuracy of 84.82%. From the standard classifiers, Deep Learning scored the highest among similar classes and the AUC scored 83.1% with an accuracy of 81.94%. The least performing model was KNN, which scored 60.3% in AUC with an F-Measure of 46.22%. It is considered as unsatisfactory classifier.

- 3. Which variables are more efficient at predicting a student's withdrawal?**

Feature weight was executed for dataset variables to find the most creditable attribute in the class label. Five metrics were used to estimate the weights. All of the attributes came in similar ranking order in all metrics. The most dominant attribute was GPA of course study.

These findings align with the outcome of reviewed research papers as shown in Table-1. This research clearly illustrates that withdrawal prediction is related in the first place to the academic performance and aligns with the findings of the researches (Aguilar et al. 2014; Almarabeh 2017; Adekitan and Noma-Osaghae 2019). Also, this research is similar to other researches (Bucos and Drăgulescu 2018; Miguéis et al. 2018; Alomari, K.M., AlHamad, A.Q. and Salloum 2019) in the outcome that the prediction classifier (Random Forest) is a particularly good classifier to predict a student's withdrawal.

Limitation

Few limitations were found during the execution of the study. First was data template which was not standardized and kept changing. Second was that the attributes title changed from one semester to another, new attributes introduced, and data formats were changed from earlier semesters. The prior led to having lots of missing data if new attributes were used, and requiring attribute mapping when a title was changed.

The third limitation was that the report in table 3 (Student Attrition Report) was not

available for the targeted study's time window. To substitute the missing data report, a feature attribute was created based on the student status condition specified in section 3.3.3.1 as a workaround solution.

Future Work

For future work, **first** is to implement the same approach on another dataset from a higher education institute which follows CHEDS template in reporting Students' information to the Ministry of Education (for example British University in Dubai). Then, to evaluate the results: Is ensemble algorithm Random Forest considered the best in predicting a student's withdrawal? **Second**, is to use a shorter time window two semesters but with a larger students' list. Then, evaluate: Do the model's results have similar outputs? **Last**, if ensemble algorithm produces similar results, then add Student Attrition Report data as mentioned in table – 3 (by this time of drafting conclusion there were 4 semesters reports available) and evaluate the result: Does the prediction classifier produce the same results?

References

- Accuracy, I.S.O. 1994. of measurement methods and results—part 1: General principles and definitions. *International Organization for Standardization, Geneva, Switzerland*
- Adekitan, A.I. and Noma-Osaghae, E. 2019. Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies* 24(2), pp. 1527–1543.
- 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. *ACM International Conference Proceeding Series* 1, pp. 103–112.
- Ahmad Tarmizi, S.S., Mutalib, S., Abdul Hamid, N.H. and Abdul Rahman, S. 2019. A Review on Student Attrition in Higher Education Using Big Data Analytics and Data Mining Techniques. *International Journal of Modern Education and Computer Science* 11(8), pp. 1–14.
- Almarabeh, H. 2017. Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science* 9(8), p. 9.
- Alomari, K.M., AlHamad, A.Q. and Salloum, S. 2019. Prediction of the digital game rating systems based on the ESRB. *Opcion* 35(19), pp. 1368–1393.
- Asif, R., Merceron, A., Ali, S.A. and Haider, N.G. 2017. Analyzing undergraduate students' performance using educational data mining. *Computers & Education* 113, pp. 177–194.
- Awad, A., Ali, A. and Gaber, T. 2020. *Mining in Educational Data: Review and Future Directions*. Springer International Publishing.
- Azevedo, A. 2019. Data mining and knowledge discovery in databases. In: *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics*. IGI Global, pp. 502–514.
- Baepler, P. and Murdoch, C.J. 2010. Academic analytics and data mining in higher education. *International Journal for the Scholarship of Teaching & Learning* 4(2)
- Bean, J.P. 1980. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education* 12(2), pp. 155–187.

- Bean, J.P. and Metzner, B.S. 1985. A conceptual model of nontraditional undergraduate student attrition. *Review of educational Research* 55(4), pp. 485–540.
- Beer, C. and Lawson, C. 2017. The problem of student attrition in higher education: An alternative perspective. *Journal of Further and Higher Education* 41(6), pp. 773–784.
- Bucos, M. and Drăgulescu, B. 2018. Predicting student success using data generated in traditional educational environments. *TEM Journal* 7(3), p. 617.
- Cogan, M.F. 2011. Predicting success of academically dismissed undergraduate students using quality point status. *Journal of College Student Retention: Research, Theory & Practice* 12(4), pp. 387–406.
- Daradoumis, T., Puig, J.M.M., Arguedas, M. and Liñan, L.C. 2019. Analyzing students' perceptions to improve the design of an automated assessment tool in online distributed programming. *Computers & Education* 128, pp. 159–170.
- Delen, D. 2010. A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems* 49(4), pp. 498–506.
- Delen, D. 2011. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory and Practice* 13(1), pp. 17–35.
- Evans, M. 2000. Planning for the transition to tertiary study: A literature review. *Journal of Institutional Research* 9(1), pp. 1–13.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Erven, G. Van 2019. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research* 94(February 2018), pp. 335–343.
- Fike, D.S. and Fike, R. 2008. Predictors of first-year student retention in the community college. *Community College Review* 36(2), pp. 68–88.
- Francis, B.K. and Babu, S.S. 2019. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *Journal of Medical Systems* 43(6)
- Hernández-Blanco, A., Herrera-Flores, B., Tomás, D. and Navarro-Colorado, B. 2019. A Systematic Review of Deep Learning Approaches to Educational Data Mining. *Complexity* 2019
- Lynch, C.F. 2017. Who prophets from big data in education? New insights and new challenges. *Theory and Research in Education* 15(3), pp. 249–271.
- Miguéis, V.L., Freitas, A., Garcia, P.J. V and Silva, A. 2018. Early segmentation of

- students according to their academic performance: A predictive modelling approach. *Decision Support Systems* 115, pp. 36–51.
- Natek, S. and Zwillling, M. 2014. Student data mining solution–knowledge management system related to higher education institutions. *Expert systems with applications* 41(14), pp. 6400–6407.
- Noble, W.S. 2006. What is a support vector machine? *Nature Biotechnology* 24(12), pp. 1565–1567.
- Pascarella, E.T. and Chapman, D.W. 1983. Validation of a theoretical model of college withdrawal: Interaction effects in a multi-institutional sample. *Research in Higher Education* 19(1), pp. 25–48.
- Del Río, C. and Insuasti, J.P. 2016. Predicting academic performance in traditional environments at higher-education institutions using data mining: A review. *Ecos de la Academia* 2(04), pp. 185–201.
- Salloum, S.A., Al-Emran, M., Abdallah, S. and Shaalan, K. 2017. Analyzing the Arab gulf newspapers using text mining techniques. In: *International Conference on Advanced Intelligent Systems and Informatics*. Springer, pp. 396–405.
- Spady, W.G. 1970. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange* 1(1), pp. 64–85.
- Spady, W.G. 1971. Dropouts from higher education: Toward an empirical model. *Interchange* 2(3), pp. 38–62.
- Tinto, V. 1975. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research* 45(1), pp. 89–125.
- Zhang, Y., Oussena, S., Clark, T. and Hyensook, K. 2010. Using data mining to improve student retention in HE: a case study. *International Conference on Enterprise Information Systems* Volume 3(12th), p. p.190-197.