5-2024

Predicting Student Attrition in Higher Education Institutions in the UAE Using Machine Learning

Dezzil M. Castelino
dmc4133@rit.edu

Predicting Student Attrition in Higher Education Institutions in the UAE Using Machine Learning

by

Dezzil M. Castelino

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs and Research

Rochester Institute of Technology (RIT Dubai)

May 2024

RIT Dubai

Master of Science in Professional Studies: Data Analytics

Graduate Thesis Approval

Student Name: Dezzil Castelino

Thesis Title: Predicting Student Attrition in Higher Education Institutions in the UAE Using Machine Learning

Graduate Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of Committee

Name: Dr. Hammou Messatfa **Date:**

Member of Committee

Acknowledgments

This Thesis would not have been possible without the guidance and help of several individuals who have contributed and provided their valuable support in the preparation and completion of the research study.

First of all, I would like to thank the almighty God for helping me to complete my thesis and master’s program successfully. I am indebted to my mentor, Dr. Hammou Messatfa whose constant guidance, support, and knowledge-sharing have enabled me to complete my thesis promptly. I thank the Department Chair of Graduate Programs and Research and all the faculty from the Data Analytics program for their dedication, and passion in delivering this program. I would also like to express my gratitude to RIT Dubai management for providing me with this opportunity to complete this master's program. I am also grateful for the dedication and consistent support of my colleagues who have been a part of this journey.

The most significant acknowledgment is reserved for my family for their constant motivation, understanding, and support during this entire master’s journey. My thesis and the master’s program would not have been complete without their endless support and encouragement.

Abstract

UAE has made significant progress in the field of education, including Higher Education, by attracting students from all around the world to various colleges and universities. Student dropout or attrition is a major issue that is faced by Higher Education Institutions (HEI). Hence, we need effective techniques to identify student data that affects attrition. This research aims to explore the use of machine learning algorithms to predict student attrition at Rochester Institute of Technology- Dubai (RIT Dubai).

In this research, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology was used as it is widely used in projects based on data mining and machine learning. Using IBM SPSS Modeler and SPSS Statistics, various machine learning models including Random Trees, Logistic Regression, Linear Support Vector Machines (LSVM), and Neural Networks were explored to find the most suitable predictive model that could accurately find student attrition and support in providing early intervention to mitigate this problem at the HEI. Logistic Regression model provided the best results with highest accuracy, AUC, precision, and recall at 1.0.

The research involves analysing student data such as demographics, results of language tests, data on academic performance, acceptance and start dates, socioeconomic factors, and other contextual data from RIT Dubai. The results of the study provide a clear understanding of the most significant predictors of student attrition that were found through machine learning. This will prove to be an important instrument for RIT Dubai to improve student retention rates.

Keywords:

Data mining, student attrition, dropout, higher education, machine learning, algorithms, random trees, logistic regression, neural networks, linear support vector machines, predictors, student retention, graduation rate, performance metrics, outliers

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1. Background

As defined by the Centre for Higher Education Data and Statistics (CHEDS), attrition refers to the students who were enrolled at the institution in the previous semester, but did not graduate and are not currently registered. Student dropout or attrition is a major issue that is faced by higher education institutions in the UAE and around the world which is leading to economic loss, student well-being, and also it has negative effects on the quality of education. This is a risk that affects all the functions of the universities and can even lead to the closure of institutions if it is not handled properly on time. With the help of supporting literature that is available on this topic, analysis of academic and non-academic factors that can affect attrition and through Machine Learning techniques this research will develop recommendations for the most suitable predictive model that can accurately find student attrition. Early intervention will support the HEIs to mitigate this problem at the right time in order to increase student retention.

1.2. Problem Statement

Rochester Institute of Technology, Dubai (RIT Dubai) is a branch campus of Rochester Institute of Technology New York. The university offers 9 undergraduate programs and 8 graduate programs. The Institutional Effectiveness Office (IE Office) at RIT Dubai is responsible for Accreditation, Quality Assurance, Assessment and Statistics. The IE Office gathers data at institutional and program levels, analyses it and submits annual reports to internal stakeholders such as the senior management and external stakeholders such as Knowledge and Human Development Authority (KHDA) and Ministry of Education (MOE). As part of the reporting process to MOE, every semester the IE office submits several datasets to CHEDS. Student Attrition and Student Enrolments are two important datasets that are submitted. The attrition dataset depends on enrolments. The problem that needs to be resolved for RIT Dubai is predicting student attrition so that senior management can take the required steps to mitigate this issue. As per the attrition datasets, RIT Dubai has approximately 9% student attrition during a semester. Both attrition and enrolment datasets have missing values that do not allow the IE Office to prepare comprehensive reports so that the senior management can find possible solutions to tackle the problem of student attrition.

1.3. Research Aim and Objectives

This research aims at exploring the use of machine learning algorithms such as LSVM, logistic regression, random trees and Neural Networks to develop an AI model for predicting student attrition in higher education institutions (HEI).

The research Objectives include:

a. Evaluate the performance of various machine learning algorithms to predict student attrition in higher education institutions.
b. Implement data cleaning and pre-processing techniques to improve the data quality.
c. Utilize feature selection techniques to find out various factors affecting students' decision to withdraw, leave or transfer from the university automatically.
d. Apply research outcomes from previous literature and methodologies used to predict student attrition and cover their limitations in this research.

1.4 Research Questions

The following research questions were addressed through this research:

- How do different ML algorithms perform in predicting student attrition in HEIs?
- What techniques can be used to enhance the data quality of the attrition and enrolment datasets?
- How can feature selection techniques assist in automating the identification of the significant predictors of student attrition at RIT Dubai?

1.5 Limitations

a) Data Limitation:

Previous research lacks data as data is limited to only one level or discipline, or a smaller number of students. In this research, the focus is on the Bachelor level and the scope for improvement is by preparing the model to focus on the Master level students list.

b) **Model Governance and Accountability:**

It is essential to have a department or staff responsible for the model governance, deployment, tracking and accountability.

c) **Monitoring and Maintenance:**

Also, it is important to monitor and maintain the model to detect and address performance degradation and concept drift.

d) **Drift Analysis:**

Detecting and adapting to changes in the data for maintaining model accuracy over time as the accuracy can change over time.

e) **Feedback Loops:**

Finally, it is also important to incorporate user feedback and environmental impact into the model's learning process for further improvement.

1.6 Structure of the Thesis

The research thesis has various chapters and all of them aim towards predicting student attrition in higher education institutions. The description per chapter is provided below:

Chapter 1: Introduction

This chapter provides the background, problem statement, research aim, objectives, research questions and the limitations of the research. The research explores the use of machine learning to predict student attrition by analysing various factors or variables that can affect student attrition in higher education institutions in the UAE.

Chapter 2: Literature Review

This chapter provides a critical evaluation of the existing research on predicting student attrition using machine learning, the research gaps that have led to comprehensive research on this topic and the key takeaways from the existing research.

Chapter 3: Research Methodology

This chapter includes the qualitative or quantitative methods used in this research. It contains a detailed discussion on the methodology.

Chapter 4: Findings and Data Analysis

This section defines the type of the dataset, its unique qualities that make them relevant to this research (visualization), data analysis, data cleaning and other measures taken to make the datasets valid, reliable and interpretable, and the problems faced during the implementation of the research study.

Chapter 5: Discussion

This chapter shows the significant results and evaluation of the findings. This section supports analytic and critical thinking on the outcomes and analysis.

Chapter 6: Conclusion

This chapter includes the summary and contributions of this research study, identified limitations and the scope for future work.

Chapter 2 – Literature Review

A prediction model for student attrition can be created with the help of machine learning, which can also provide early notice to institutional authorities so they can take required action with students who are at risk of dropping out (Del Bonifro, Gabbrielli, Lisanti, & Zingaro, 2020). Predictive Modelling is a set of techniques that includes building and using models that can make predictions based on patterns that are extracted from data (Kelleher, Namee, & D'arcy, 2020). Several articles have been published about student attrition, the factors leading towards it that ultimately result into students' drop out and the possible ML techniques that can be used to tackle this issue. This literature review highlights the existing literature about this topic, what are the possible ML techniques that have been successfully used to resolve this issue and what are the limitations in the existing studies.

Student Attrition not only affects the student who drops out but also affects the university, society as well as the country. (Barefoot, 2007) states that depending on their degree of maturity, preparedness for college, or sentiments of personal belonging in the college, dropout from higher education has varying effects on students. Student attrition has serious effects on the institution. Small private institutions are often fully dependent on the tuition fee and the loss of students can have a disastrous effect on their operating budget.

According to (Latif, Choudhary, & Hammayun, 2015), education is very important for economic development of the countries as it promotes innovation, entrepreneurship, productivity; provides higher employment opportunities, and also supports women's empowerment. According to UNICEF 80 million students drop out before the completion of their elementary education in India due to financial reasons, health reasons, family issues, poverty or gender inequality. Dropout leads to a lower literacy rate and in turn, leads to unemployment or jobs with lower wages and reliance on government assistance. (Catterall, 1985) pointed out that every year, student dropout costs countries over $200 billion during their lifetime due to lost earnings and unrealized tax revenue. The number of student dropouts affects the future economic development and overall development of the countries. According to the (OECD, 2022) report- Brazil, Greece, Italy, and South Africa have the highest share of young people suffering long-term unemployment: around 5% or more of 18–24-year-olds in these countries were not in education and had been unemployed for at least 12 months in the first quarter of 2021 which leaves them, particularly at risk of long-term detachment from the labour market. (Helbling & Sacchi, 2014) and (Bäckman &

Nilsson, 2016) indicate that the youth who are neither employed nor in formal education or training (NEET) not only miss out on immediate learning and employment opportunities, but also suffer from long-term effects and this status has been associated with various adverse outcomes, such as lower employment rates and lower earnings, poor mental health and social exclusion.

(Zang & Rangwala, 2018) explained that universities and colleges must develop data-driven AI systems to identify students at risk earlier and provide timely guidance and support for them. Further, the authors confirmed that most of the current classification approaches on early dropout prediction are unable to utilize all the information from historical data from previous cohorts to predict dropouts of current students in a few semesters. The iterative Logistic Regression (ILR) method was used and this framework was able to make full use of historical student data and effectively predict students at risk of failing or dropping out in future semesters.

(Niyogisubizo, Liao, Nziyumva, Murwanashyaka, & Nshimyumukiza, 2022) stated that student attrition is a global issue which not only affects the individual student who drops out but also affects the school, family and society. Big data is the most important technology used in data analysis which has been used mostly in the past research on student attrition. However, in this research the authors propose a novel stacking ensemble based on a hybrid of Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), and Feed-forward Neural Networks (FNN) to predict student's dropout in university classes. This method showed greater performance compared to the base models using accuracy and area under the curve (AUC). The research shows that based on influential factors, students who are at risk of dropping out of school can be identified. Various education representatives can use this information to identify uncontrolled behavior that may increase the likelihood of dropping out can take proactive measures before problems arise to prevent it.

(Ahmad Tarmizi, Mutalib, Abdul Hamid, & Abdul Rahman, 2019) further elaborated that students can be considered as dropouts when they withdraw from a course by cancelling their enrolment in the program, fail to continue in the following semester, accept the offer but do not register or enrol into the program, miss classes and fail in the examination several times. The factors that lead to student attrition can be divided into three categories namely, student information, academic information and family information. As part of these categories, selective attributes were found to be most effective on the student's attrition decision. These attributes include gender, program code (defining the student major), age, GPA, family income and loan, race, parents' education, parents' occupation, high school grade and marital status. The authors of the research conclude that when institutions use cutting-edge analytical methods like big data analytics and data mining techniques in their management process, attrition problems can be resolved.

(Mnyawami, Maziku, & Mushi, 2022) focused on the ongoing dropout issue in secondary schools. According to this research dropout issue in secondary schools is particularly because of improper identification of the root cause and the lack of formal procedures that may be used to estimate the severity of the issue. In this research, the authors used AutoML model to increase prediction accuracy by choosing the appropriate hyper-parameters, features, and ML algorithm for the obtained dataset. The suggested model attained high prediction accuracy of over 90% which indicates that the selection of the right factors that contribute to student dropouts is the key to monitor student attrition in order to apply early intervention to mitigate it.

(Delen, 2010) further stated that student retention is the most important priority for decision makers in higher education institutions and to improve retention we need to have a good understanding of the factors affecting attrition. In this research the author used several data mining techniques, both individual and ensembles, to develop analytical models to predict and explain the reasons behind freshmen student attrition. Four classification methods- artificial neural networks, decision trees, support vector machines and logistic regression and three ensemble techniques- bagging, boosting and information fusion were used. Ensembles provided better results compared to individual models and balanced dataset provided better prediction results compared to unbalanced dataset. Educational and financial variables were the most important factors affecting attrition. The limitation of this study is that it is limited to predict attrition based on only institutional data.

(Chung & Lee, 2019) concentrated on developing machine learning (ML) prediction models to resolve secondary school dropout as they have a great potential to develop early warning systems to recognize students at risk of dropping out in advance and to assist them. Predictive analytics build models to predict based on past data and ML is used to train the models. The authors used random forests ML model to predict students' dropouts and the model showed excellent performance with 95% accuracy and 97% AUC. The research limitations include limiting the research to descriptive features and potential inaccuracy in the weights used in calculating the features of the model.

(Jadric, Garaca, & Cukusic, 2010) indicated that the majority of the dropouts were in their first years of studies. Hence, the management of the HEIs should work on better planning, management and controlling educational processes to enhance the effectiveness of studying. The authors applied Data Mining methods to predict student dropout or attrition. Methods such as logistic regression, decision trees, and neural networks were used and among them decision trees demonstrated greater ability to predict student attrition accurately. The research was limited to the Faculty of Economics and the data obtained was only for first-year courses.

Further, (Kurian & Al-Assaf, 2020) conducted research to identify the influence of high school curriculum on student performance in universities. The research study's academic and non-cognitive measures of the students. Linear regression, decision tree and random forest were used for prediction. Moreover, a survey was conducted to collect non-cognitive skills. It was found that students from different curriculum performed differently at the university. The study was limited to only undergraduate students at RIT- Dubai. So, scaling it to bigger population can provide better insights and more reliable results.

(Kemper, 2020) proposed models based on examination data. Two most suitable ML techniques were used for predicting student dropout- decision trees and logistic regression as both the methods provided high prediction accuracies of more than 83%. Both methods provided prediction accuracies of up to 95% after three semesters.

(Willcoxson, Cotter, & Joy, 2011) focused on student experience in subsequent years of the university as compared to several studies related to student experience in freshmen years. Factors affecting the intention to dropout were differentiated by year of study and then they were differentiated by the university attended. The results of this research showed that first-year student attrition or retention is substantially influenced by the teaching staff's approachability and capacity to make courses engaging and challenging. The second-year students may intend to drop out due to lack of their own ability to succeed. However, confidence-building or skill-building activities can help them to stay. Similarly, the third-year students who are lacking the ability and unsure of their career direction are more likely to leave the university. Such students can be retained by offering further study options which can provide them with commitment building alternatives. Also, connection with alumni, reiterating graduate employment outcomes etc., can help retain such students. This study is limited to university experience of only business students. The attrition factors may be different for students from other disciplines.

(Barramuño, Meza-Narváez, & Gálvez-García, 2022) used ML to generate predictive methods to identify attrition from a database of 336 university students taking upper- year courses. The study showed that the best accuracy of 86.3% was gained using Subspace KNN algorithm and the classifier – RUSboosted trees provided lowest false negatives and higher sensitivity of the algorithms used at 78% with specificity of 86%. While the predictive method was successful in identifying student attrition in the university program, the study was limited to data from only upper-level courses.

Key Takeaways

- ML models can predict student attrition in higher education institutions.
- There is effectiveness and precision in predicting student attrition using ML models.
- Feature selection techniques can successfully identify and predict the significant factors affecting attrition with high performance.
- Several academic/non-academic factors affecting student attrition can be analysed using ML models to find any existing patterns which can be utilized to create early warning systems, enabling institutions to take necessary actions and support at risk students as needed.
- Effective data cleaning and pre-processing techniques can improve the data quality.
- Researchers have worked on finding the factors leading to student attrition by using analytical, cognitive, and ensemble methods.
- Past research limitations include lack of data, data limited to only one level or discipline, or a smaller number of students.

Chapter 3- Research Methodology

The research uses CRISP-DM (Cross Industry Standard Process for Data Mining) as the methodology framework. It is a widely adopted methodology for data mining and ML-related projects. As shown in figure 1 below, CRISP-DM offers an organized methodology for planning as well as executing data mining projects. This structure enables projects to efficiently execute faster. The various stages of CRISP-DM can be used for several research methodologies. Using CRISP-DM allows research to be structured, ensuring the right data is collected and analysed facilitating data-driven decision-making (Martinez-Plumed, et al., 2021). Since it follows a standardized process, it provides assurance to researchers that their methods are transparent and can be repeated. This helps in reducing failures, saves time and other resources while increasing high-quality, research-driven and reliable results.

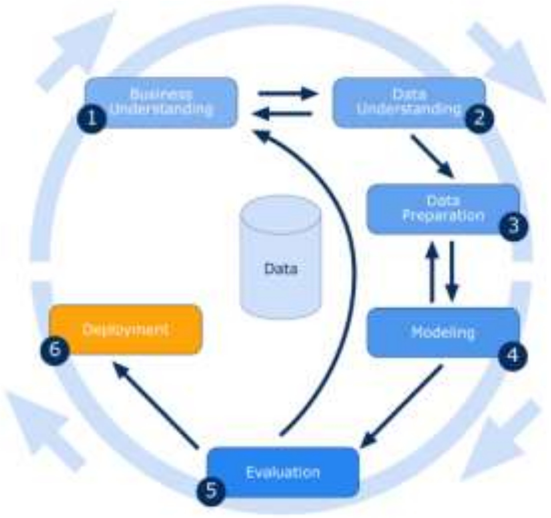


Figure 1-CRISP-DM Process

(Source: linkedin.com-introduction-crisp-dm-framework-data-science)

Machine learning is a set of algorithms that can learn from a dataset (Goodfellow, Bengio, & Courville, 2016). It is a branch of Artificial intelligence (AI) that focuses on using data and algorithms to simulate people's learning along with a gradual increase in the accuracy of the model. Algorithms are trained using statistical techniques to produce classifications or predictions and to find important insights in data mining projects. The decisions made as a result of these insights influence growth indicators in organizations.

The ML process is described in figure 2 below. The process includes collecting the required data, feature selection and data preparation, choosing the ML algorithm to create the model, training the model, testing or evaluating the model and finally the model is used for making predictions.

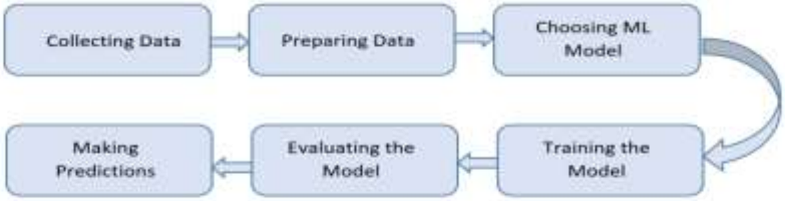


Figure 2- Machine Learning Process

Machine learning has two categories: supervised and unsupervised learning. In Supervised Learning, ML algorithms learn the relationship between descriptive features which are called predictors, and target features that is the outcomes in a dataset (Chung & Lee, 2019). Using the trained model from supervised learning, we can predict the outcome of future observations or better understand the relationship between the outcome and predictors accurately. For instance, supervised learning can be used to train machine learning algorithms to understand the correlation between different predictors and student attritors. A dataset must contain both the target feature or outcome and descriptive features or predictors in Supervised Learning (Chung & Lee, 2019). So, the dataset for supervised learning is called a labelled dataset as the dataset contains a target or a label that supervises the learning process. In supervised learning, statistical models such as Logistic regression, linear regression, and support vector machines are used (James, Witten, Hastie, & Tibshirani, 2013). In this thesis, supervised learning technique was applied for training and testing the data sets.

Chapter 4- Findings and Data Analysis

4.1 Data Description

This chapter provides a detailed analysis of the two datasets used in the research. It includes descriptive statistics, exploratory data analysis, visualization of key features, and statistical analysis. Further, it discusses about feature importance analysis to identify the most significant factors for predicting attrition.

4.1.1. Data Collection

As discussed in Chapter 1, this research involves quantitative data that is student data from RIT Dubai. The Institutional Research Ethics Board (IREB) at RIT Dubai was requested to approve the usage of the data that was available within the Institutional Effectiveness (IE) office. After IREB approval, the IE office pre-processed the data to remove any personal identifiers from the datasets.

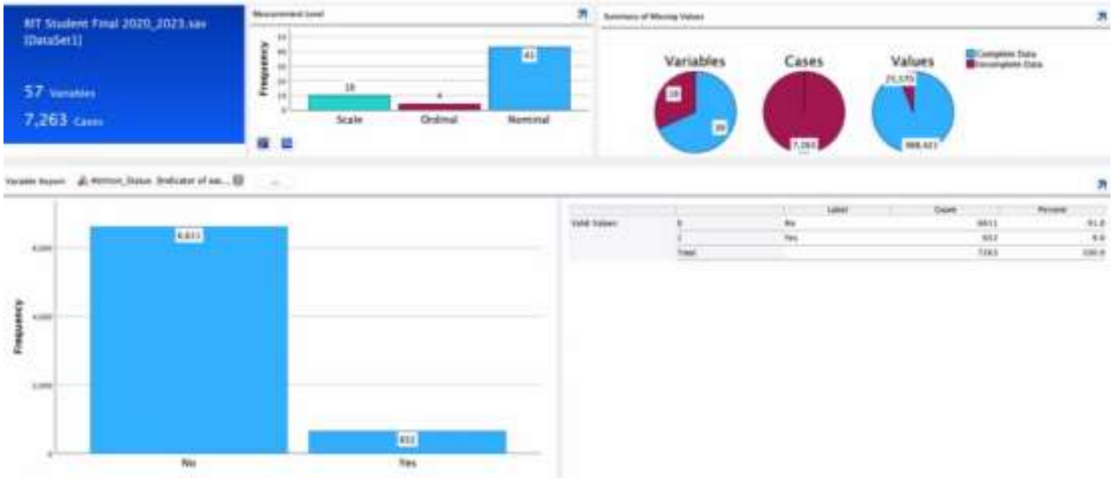
The two datasets used for this research- Enrolments and Attrition, were then released by the IE Office to conduct the research to predict student attrition. The selected datasets are from three academic years: 2020-2021 until 2022-2023. The datasets contain student information from both undergraduate and graduate level. The enrolment dataset has 50 variables and attrition dataset has 10 variables. The objects or cases include data on student demographics, English test scores, Program major and specialization, acceptance date, mode of study, number of credits, GPA, last completed high school or university information etc.

IBM SPSS Modeler and SPSS Statistics were used for data exploration for understanding and analysing the data and also to find the relationship between the datasets and the variables, finding null values, and for gathering clear insights about the attributes. Further, visualization is used to display and summarize the findings.

4.1.2. Data Profile

As part of data preparation, the data in the two datasets was merged into 1 comprehensive dataset and the duplicate variables were removed. So, we have 57 unique variables and 7263 cases as shown in the figure 1 below. Out of the 57 variables, 43 were nominal (nationality), 4 were ordinal (last completed HS/ BS) and 10 were categorical (Gender). 39 variables have completed cases or objects and 18 variables have incomplete cases or objects.

The two datasets (Enrolment and Attrition) are labelled datasets that include ‘Attrition’ as the target feature and descriptive features such as Program Code, Student Major, Language Test Proficiency Exam Score, High School Exit Score, Registered Credits, Overall GPA, etc. Figure 3 below also shows the attrition status:


RIT Student Final 2020_2023.sav [DataSet1]

57 Variables

7,263 Cases

		Label	Count	Percent
Valid Values	0	No	6611	91.0
	1	Yes	652	9.0
	Total		7263	100.0

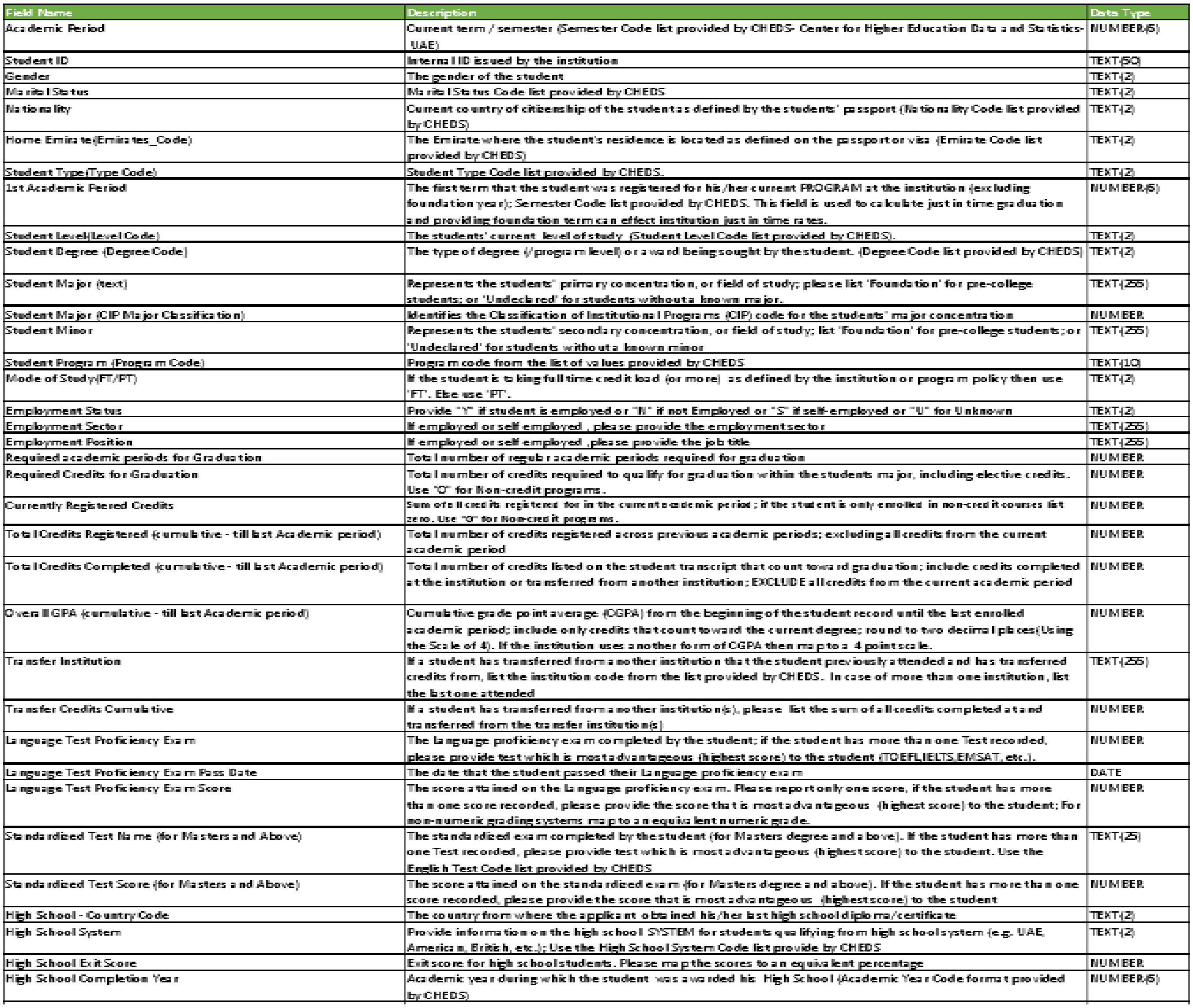
Figure 3- Data Description

Characteristics or attributes that may be measured, adjusted, or controlled are referred to as *Variables* and a *Target Variable* is the main feature that will provide us with a deeper understanding of the dataset. As described in the figure above, 0 and 1 shows the status of Attrition. Data shows 652 cases or 9% Attrition in the available datasets.

4.1.3. Data Dictionary

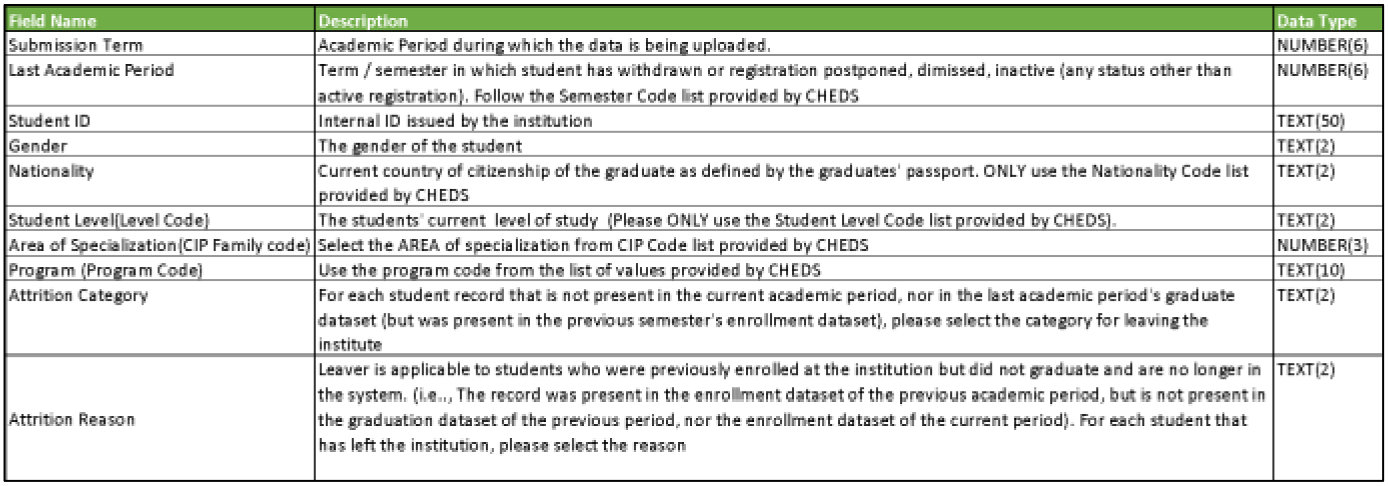
The RIT Dubai Enrolments dataset has 50 variables or attributes and the attrition dataset has 10 variables or attributes.

Table 1-Data Dictionary- Enrollments



Field Name	Description	Data Type
Academic Period	Current term / semester (Semester Code list provided by CHEDS- Center for Higher Education Data and Statistics- UAE)	NUMBER(6)
Student ID	Internal ID issued by the institution	TEXT(50)
Gender	The gender of the student	TEXT(2)
Marital Status	Marital Status Code list provided by CHEDS	TEXT(2)
Nationality	Current country of citizenship of the student as defined by the students' passport (Nationality Code list provided by CHEDS)	TEXT(2)
Home Emirate(Emirates_Code)	The Emirate where the student's residence is located as defined on the passport or visa (Emirate Code list provided by CHEDS)	TEXT(2)
Student Type(Type Code)	Student Type Code list provided by CHEDS.	TEXT(2)
1st Academic Period	The first term that the student was registered for his/her current PROGRAM at the institution (excluding foundation year); Semester Code list provided by CHEDS. This field is used to calculate just in time graduation and providing foundation term can effect institution just in time rates.	NUMBER(6)
Student Level(Level Code)	The students' current level of study (Student Level Code list provided by CHEDS).	TEXT(2)
Student Degree (Degree Code)	The type of degree (/program level) or award being sought by the student. (Degree Code list provided by CHEDS)	TEXT(2)
Student Major (text)	Represents the students' primary concentration, or field of study; please list 'Foundation' for pre-college students; or 'Undeclared' for students without a known major.	TEXT(255)
Student Major (CIP Major Classification)	Identifies the Classification of Institutional Programs (CIP) code for the students' major concentration	NUMBER
Student Minor	Represents the students' secondary concentration, or field of study; list 'Foundation' for pre-college students; or 'Undeclared' for students without a known minor	TEXT(255)
Student Program (Program Code)	Program code from the list of values provided by CHEDS	TEXT(10)
Mode of Study(FT/PT)	If the student is taking full time credit load (or more) as defined by the institution or program policy then use 'FT'. Else use 'PT'.	TEXT(2)
Employment Status	Provide "Y" if student is employed or "N" if not Employed or "S" if self-employed or "U" for Unknown	TEXT(2)
Employment Sector	If employed or self employed , please provide the employment sector	TEXT(255)
Employment Position	If employed or self employed ,please provide the job title	TEXT(255)
Required academic periods for Graduation	Total number of regular academic periods required for graduation	NUMBER
Required Credits for Graduation	Total number of credits required to qualify for graduation within the students major, including elective credits. Use "0" for Non-credit programs.	NUMBER
Currently Registered Credits	Sum of all credits registered for in the current academic period ; if the student is only enrolled in non-credit courses list zero. Use "0" for Non-credit programs.	NUMBER
Total Credits Registered (cumulative - till last Academic period)	Total number of credits registered across previous academic periods; excluding all credits from the current academic period	NUMBER
Total Credits Completed (cumulative - till last Academic period)	Total number of credits listed on the student transcript that count toward graduation; include credits completed at the institution or transferred from another institution; EXCLUDE all credits from the current academic period	NUMBER
Overall GPA (cumulative - till last Academic period)	Cumulative grade point average (CGPA) from the beginning of the student record until the last enrolled academic period; include only credits that count toward the current degree; round to two decimal places(Using the Scale of 4). If the institution uses another form of CGPA then map to a 4 point scale.	NUMBER
Transfer Institution	If a student has transferred from another institution that the student previously attended and has transferred credits from, list the institution code from the list provided by CHEDS. In case of more than one institution, list the last one attended	TEXT(255)
Transfer Credits Cumulative	If a student has transferred from another institution(s), please list the sum of all credits completed at and transferred from the transfer institution(s)	NUMBER
Language Test Proficiency Exam	The language proficiency exam completed by the student; if the student has more than one Test recorded, please provide test which is most advantageous (highest score) to the student (TOEFL,IELTS,EMSAT, etc.).	NUMBER
Language Test Proficiency Exam Pass Date	The date that the student passed their Language proficiency exam	DATE
Language Test Proficiency Exam Score	The score attained on the language proficiency exam. Please report only one score, if the student has more than one score recorded, please provide the score that is most advantageous (highest score) to the student; For non-numeric grading systems map to an equivalent numeric grade.	NUMBER
Standardized Test Name (for Masters and Above)	The standardized exam completed by the student (for Masters degree and above). If the student has more than one Test recorded, please provide test which is most advantageous (highest score) to the student. Use the English Test Code list provided by CHEDS	TEXT(25)
Standardized Test Score (for Masters and Above)	The score attained on the standardized exam (for Masters degree and above). If the student has more than one score recorded, please provide the score that is most advantageous (highest score) to the student	NUMBER
High School - Country Code	The country from where the applicant obtained his/her last high school diploma/certificate	TEXT(2)
High School System	Provide information on the high school SYSTEM for students qualifying from high school system (e.g. UAE, American, British, etc.); Use the High School System Code list provide by CHEDS	TEXT(2)
High School Exit Score	Exit score for high school students. Please map the scores to an equivalent percentage	NUMBER
High School Completion Year	Academic year during which the student was awarded his High School (Academic Year Code format provided by CHEDS)	NUMBER(6)

Table 2-Data Dictionary- Attrition



Field Name	Description	Data Type
Submission Term	Academic Period during which the data is being uploaded.	NUMBER(6)
Last Academic Period	Term / semester in which student has withdrawn or registration postponed, dimissed, inactive (any status other than active registration). Follow the Semester Code list provided by CHEDS	NUMBER(6)
Student ID	Internal ID issued by the institution	TEXT(50)
Gender	The gender of the student	TEXT(2)
Nationality	Current country of citizenship of the graduate as defined by the graduates' passport. ONLY use the Nationality Code list provided by CHEDS	TEXT(2)
Student Level(Level Code)	The students' current level of study (Please ONLY use the Student Level Code list provided by CHEDS).	TEXT(2)
Area of Specialization(CIP Family code)	Select the AREA of specialization from CIP Code list provided by CHEDS	NUMBER(3)
Program (Program Code)	Use the program code from the list of values provided by CHEDS	TEXT(10)
Attrition Category	For each student record that is not present in the current academic period, nor in the last academic period's graduate dataset (but was present in the previous semester's enrollment dataset), please select the category for leaving the institute	TEXT(2)
Attrition Reason	Leaver is applicable to students who were previously enrolled at the institution but did not graduate and are no longer in the system. (i.e.., The record was present in the enrollment dataset of the previous academic period, but is not present in the graduation dataset of the previous period, nor the enrollment dataset of the current period). For each student that has left the institution, please select the reason	TEXT(2)

4.2 Exploratory Data Analysis

4.2.1. Statistics Summary

As shown in table 3 below, the continuous variables have the values for min, max, mean, standard deviation and skewness. However, the categorical variables have a number of unique values.

Table 3- Statistics Summary

Field	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique
RequiredacademicperiodsforGraduation	Continuous	0	99	8.31	6.549	12.276	--
RequiredCreditsforGraduation	Continuous	0	129	109.564	36.44	-1.716	--
CurrentlyRegisteredCredits	Continuous	0	22	12.559	4.7	-1.031	--
TotalCreditsRegisteredcumulativetilllastAcademicperiod	Continuous	0	214	42.956	41.95	0.895	--
TotalCreditsCompletedcumulativetilllastAcademicperiod	Continuous	0	163	38.173	36.753	0.796	--
OverallGPAcumulativetilllastAcademicperiod	Continuous	0	4	2.25	1.354	-0.647	--
TransferCreditsCumulative	Continuous	0	113	1.518	7.597	6.863	--
LanguageTestProficiencyExamPassDate	Continuous	1899-12-31	1/1/9999	--	--	--	--
LanguageTestProficiencyExamScore	Continuous	0	2000	408.852	656.498	1.329	--
HighSchoolExitScore	Continuous	0	100	77.771	27.637	-2.116	--
HighSchoolCompletionYear	Continuous	199495	999999	443803.412	366822.382	0.857	--

Attrition	Flag	0	1	--	--	--	2
Gender	Nominal	--	--	--	--	--	3
Nationality	Nominal	--	--	--	--	--	89
StudentLevelLevelCode	Nominal	--	--	--	--	--	7
MaritalStatus	Nominal	--	--	--	--	--	4
HomeEmirateEmirates_Code	Nominal	--	--	--	--	--	9
StudentTypeTypeCode	Nominal	--	--	--	--	--	2
StudentDegreeDegreeCode	Nominal	--	--	--	--	--	4
StudentMajortext	Nominal	--	--	--	--	--	22
StudentProgramProgramCode	Nominal	--	--	--	--	--	20
ModeofStudyFTPT	Nominal	--	--	--	--	--	2
EmploymentStatus	Nominal	--	--	--	--	--	3
LanguageTestProficiencyExam	Nominal	7	30	--	--	--	10
HighSchoolCountryCode	Nominal	--	--	--	--	--	51
HighSchoolSystem	Nominal	--	--	--	--	--	22
HighSchoolEquivalencyIndicator	Nominal	--	--	--	--	--	6
@12thGradestreamStreamCode	Nominal	--	--	--	--	--	6
LastCompletedHigherEducationDegree	Nominal	--	--	--	--	--	4
OutgoingExchangestudentIndicator	Nominal	--	--	--	--	--	2
AcademicPeriod	Ordinal	202001	202203	--	--	--	6
@1stAcademicPeriod	Ordinal	201101	202203	--	--	--	25
StudentMajorCIPMajorClassification	Ordinal	4.03	100	--	--	--	17

4.2.2. Data Cleaning

In the below sections, we have identified and handled missing data, outliers, or inconsistencies in the dataset.

4.2.2.1. Techniques for Handling Missing Data

According to (Graham, 2012), missing data can be of three types as described below:

1. Missing completely at random (MCAR)
   Data is said to be missing completely at random if the probability of missing is the same for all cases. This indicates that the reasons for missing the data in not depend on the values of the data.

2. Missing at random (MAR)

   Data is said to be missing at random (MAR) if the probability of missing is the same only within groups defined by the observed data. In case of MAR, we must take the causes of missingness into account by including those variables in the analysis model to avoid estimation bias.

3. Missing not at random (MNAR)

   In the case of MNAR, the cause of missingness is not measured and is not available for analysis.

   Missing data can be addressed by removing the variables, removing the cases or by imputing the missing values. In this research, the variables which had a large number of null values are removed as they did not affect any other variable, mainly attrition. For example, in this research, the variables High School Equivalency Number, Area of Specialization, Student Minor, Employment Sector, Employment Position etc., were removed as they do not affect attrition.

(Chauvet, Deville, & & Haziza, 2011) describe the different data Imputation methods which include the following:

1. *Fixed* where the same value is used for all cases.
2. *Random* where different or random values or cases are used based on a normal or uniform distribution which allows variation in the field with imputed values.
3. *Expression* which allows to create our own equation to specify missing values.
4. *Algorithm* which uses a value predicted by a classification and regression tree model.

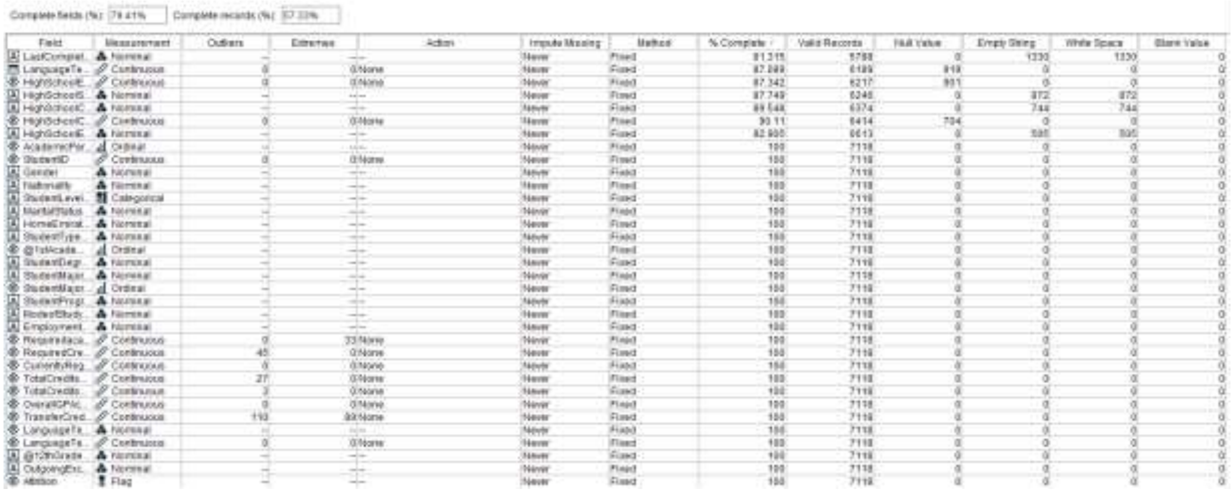
In this research study, *Expression* and *Algorithm* methods were used to add the missing values. The missing values in the variables were imputed as described below:

1. The missing values in Required Academic Periods for Graduation were replaced with 8.310 (Mean value).
2. The missing values in Required Credits for Graduation were replaced with 109.564 (Mean value).
3. The missing values in Currently Registered Credits were replaced with 12.559 (Mean value).
4. The missing values in Total Credits Registered Cumulative till last academic period were replaced with 42.956 (mean value).

5. The missing values in Total Credits Completed Cumulative till last academic period were replaced with 38.173 (mean value).
6. The missing values in Overall GPA Cumulative till last academic period were replaced with 2.250 (mean value).
7. The missing values in Transfer Credits Cumulative were replaced with 1.518 (mean value).
8. The missing values in Student Level Code were replaced with FR (mean value).
9. The missing values in Language Test Proficiency Exam was replaced with an algorithm '$R-LanguageTestProficiencyExam'.
10. The missing values in Language Test Proficiency Exam Score was replaced with an algorithm: '$R-LanguageTestProficiencyExamScore'.

The results of the imputation are as follows:

Table 4-Data Quality after Imputation



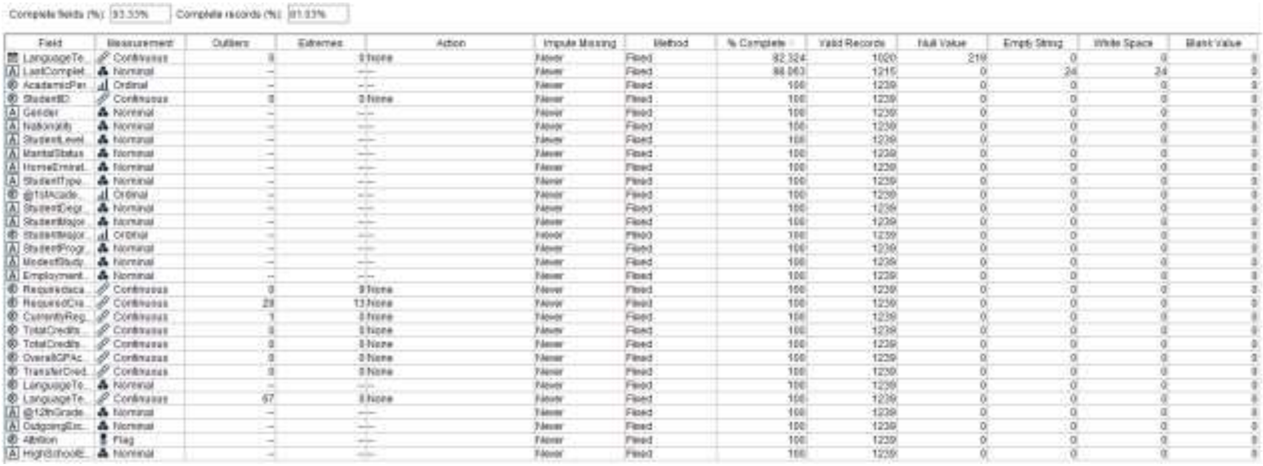
Complete fields (%): 79.41%   Complete records (%): 57.33%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
LastComplet...	Nominal	--	--	--	Never	Fixed	81.315	5788	0	1330	1330	0
LanguageTe...	Continuous	0	0	None	Never	Fixed	87.289	6213	905	0	0	0
HighSchoolE...	Continuous	0	0	None	Never	Fixed	87.342	6217	901	0	0	0
HighSchoolS...	Nominal	--	--	--	Never	Fixed	87.749	6246	0	872	872	0
HighSchoolC...	Nominal	--	--	--	Never	Fixed	89.548	6374	0	744	744	0
HighSchoolC...	Continuous	0	0	None	Never	Fixed	90.11	6414	704	0	0	0
HighSchoolE...	Nominal	--	--	--	Never	Fixed	92.905	6613	0	505	505	0
AcademicPer...	Ordinal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentID	Continuous	0	0	None	Never	Fixed	100	7118	0	0	0	0
Gender	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
Nationality	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentLevel...	Categorical	--	--	--	Never	Fixed	100	7118	0	0	0	0
MaritalStatus	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
HomeEmirat...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentType...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
@1stAcade...	Ordinal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentDegr...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentMajor...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentMajor...	Ordinal	--	--	--	Never	Fixed	100	7118	0	0	0	0
StudentProgr...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
ModeofStudy...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
Employment...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
RequiredAca...	Continuous	0	33	None	Never	Fixed	100	7118	0	0	0	0
RequiredCre...	Continuous	45	0	None	Never	Fixed	100	7118	0	0	0	0
CurrentlyReg...	Continuous	0	0	None	Never	Fixed	100	7118	0	0	0	0
TotalCredits...	Continuous	27	0	None	Never	Fixed	100	7118	0	0	0	0
TotalCredits...	Continuous	3	0	None	Never	Fixed	100	7118	0	0	0	0
OverallGPAc...	Continuous	0	0	None	Never	Fixed	100	7118	0	0	0	0
TransferCred...	Continuous	110	89	None	Never	Fixed	100	7118	0	0	0	0
LanguageTe...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
LanguageTe...	Continuous	0	0	None	Never	Fixed	100	7118	0	0	0	0
@12thGrade...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
OutgoingExc...	Nominal	--	--	--	Never	Fixed	100	7118	0	0	0	0
Attrition	Flag	--	--	--	Never	Fixed	100	7118	0	0	0	0

With the help of Imputation, the data quality was improved from 50% to 79% with 57% completed records (data audit with 34 variables). So, we reduced the incomplete variables from17 to 7 which are Lastcompletedhighereducationdegree, Languagetestproficiencyexampassdate, Highschoolexitscore, Highschoolsystem, Highschoolcountrycode, Highschoolcompletionyear and Highschoolequivalencyindicator.

Further the data quality was improved to 93% with 81% completed records (data audit with 34 variables) as shown in the table below. The missing values in the five variables: Highschoolexitscore, Highschoolsystem, Highschoolcountrycode, Highschoolcompletionyear and Highschoolequivalencyindicator were recorded as unknown.

Table 5-Data Quality Improved



Complete fields (%): 93.33% Complete records (%): 81.83%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
LanguageTe...	Continuous	0	0	None	Never	Fixed	92.324	1020	219	0	0	0
LastComplet...	Nominal	–	–	–	Never	Fixed	98.063	1215	0	24	24	0
AcademicPer...	Ordinal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentID	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
Gender	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
Nationality	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentLevel	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
MaritalStatus	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
HomeEmirat...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentType...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
@1stAcade...	Ordinal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentDegr...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentMajor...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentMajor...	Ordinal	–	–	–	Never	Fixed	100	1239	0	0	0	0
StudentProgr...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
ModeofStudy	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
Employment...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
Requirededuca...	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
RequiredCre...	Continuous	23	13	None	Never	Fixed	100	1239	0	0	0	0
CurrentlyReg...	Continuous	1	0	None	Never	Fixed	100	1239	0	0	0	0
TotalCredits...	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
TotalCredits...	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
OverallGPAc...	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
TransferCred...	Continuous	0	0	None	Never	Fixed	100	1239	0	0	0	0
LanguageTe...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
LanguageTe...	Continuous	67	0	None	Never	Fixed	100	1239	0	0	0	0
@12thGrade...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
OutgoingExc...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0
Attrition	Flag	–	–	–	Never	Fixed	100	1239	0	0	0	0
HighSchoolE...	Nominal	–	–	–	Never	Fixed	100	1239	0	0	0	0

2 variables- Lastcompletedhighereducationdegree and Languagetestproficiencyexampassdate were incomplete and these were not at random. Hence, they were not imputed. So, this was used as the final dataset containing 28 variables with 1239 students and the student body consisted of only Bachelor level data.

4.2.2.2. Approaches for Detecting Outliers

An outlier is a response or observation that lies at an unusual distance from other observations or responses (Stehlik-Barry & Babinec, 2017). This subsection explains the approaches or methods that are applied to detect outliers. These techniques include:

- Statistical Method: Using statistical measures such as the mean and standard deviation to identify observations that fall outside a predefined threshold.
- Machine learning-based anomaly detection technique: Leveraging machine learning algorithms to identify anomalies based on deviations from the normal behavior of the dataset.

In this study, the Statistical approach was applied. Table 6 below shows the fields containing outliers or extreme outliers. Every value of the field that is greater than the mean + 3 Standard deviations is an outlier. Similarly, every value that is less than the mean - 3 Standard deviations is an outlier. For example, **TotalCreditsRegisteredcumulativetilllastAcademicperiod** has 27 outliers and **TransferCreditsCumulative** has 110 outliers. Every value of the field that is greater than the mean + 5 Standard deviations is an extreme outlier. Similarly, every value that is less than the mean - 5 Standard deviations is an extreme outlier. For example, **RequiredacademicperiodsforGraduation** has 33 extreme outliers and **TransferCreditsCumulative** has 89 extreme outliers.

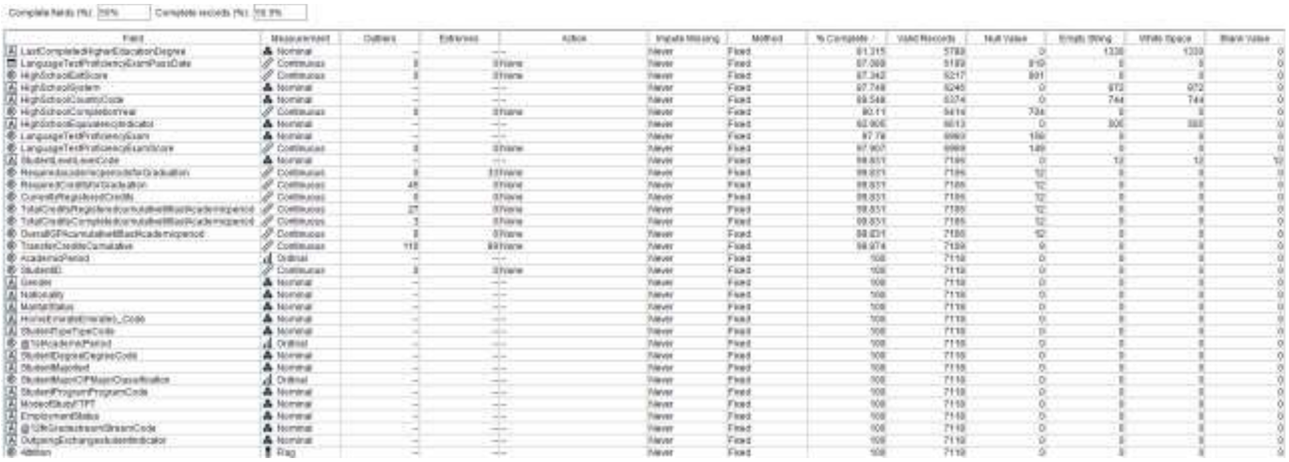
Table 6- Outliers and Extreme Outliers

Field	Measurement	Outliers	Extremes
RequiredacademicperiodsforGraduation	Continuous	0	33
TotalCreditsCompletedcumulativetilllastAcademicperiod	Continuous	3	0
TotalCreditsRegisteredcumulativetilllastAcademicperiod	Continuous	27	0
RequiredCreditsforGraduation	Continuous	45	0
TransferCreditsCumulative	Continuous	110	89

4.2.2.3. Data Quality

The data underwent cleaning or pre-processing which includes removal of duplicate records. The data in the two datasets was merged into 1 dataset and the duplicate variables were removed. So, 57 unique variables were finalized. As shown in Figure 3, initially there were **25,570** missing values. Table 7 below shows the data quality at the beginning of the research, where the completed fields were only at 50% and completed records were 56.9%. For example, the variable “Lastcompletedhighereducationdegree” had 81.3% completion rate with 1330 missing values. Similarly, the variable “Highschoolexitscore” had 87.3% completion rate. However, this variable had 901 null values.

Table 7-Data Quality



Complete fields (%): 59% Complete records (%): 78.9%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
LastCompletedHigherEducationDegree	Nominal	--	--		Never	Fixed	81.315	5789	0	1330	1330	0
LanguageTestProficiencyExamPassDate	Continuous	0	0 None		Never	Fixed	87.388	6189	910	0	0	0
HighSchoolExitScore	Continuous	0	0 None		Never	Fixed	87.342	6217	901	0	0	0
HighSchoolSystem	Nominal	--	--		Never	Fixed	87.748	6246	0	872	872	0
HighSchoolCountryCode	Nominal	--	--		Never	Fixed	89.548	6374	0	744	744	0
HighSchoolCompletionYear	Continuous	0	0 None		Never	Fixed	90.11	6414	734	0	0	0
HighSchoolEquivalencyIndicator	Nominal	--	--		Never	Fixed	92.906	6613	0	506	506	0
LanguageTestProficiencyExam	Nominal	--	--		Never	Fixed	97.78	6960	159	0	0	0
LanguageTestProficiencyExamScore	Continuous	0	0 None		Never	Fixed	97.907	6969	149	0	0	0
StudentLevelLevelCode	Nominal	--	--		Never	Fixed	99.831	7106	0	12	12	12
RequiredacademicperiodsforGraduation	Continuous	0	11 None		Never	Fixed	99.831	7106	12	0	0	0
RequiredCreditsforGraduation	Continuous	48	0 None		Never	Fixed	99.831	7106	12	0	0	0
CurrentlyRegisteredCredits	Continuous	0	0 None		Never	Fixed	99.831	7106	12	0	0	0
TotalCreditsRegisteredcumulativetilllastAcademicperiod	Continuous	27	0 None		Never	Fixed	99.831	7106	12	0	0	0
TotalCreditsCompletedcumulativetilllastAcademicperiod	Continuous	3	0 None		Never	Fixed	99.831	7106	12	0	0	0
OverallGPAcumulativetilllastAcademicperiod	Continuous	0	0 None		Never	Fixed	99.831	7106	12	0	0	0
TransferCreditsCumulative	Continuous	112	89 None		Never	Fixed	99.874	7109	9	0	0	0
AcademicPeriod	Ordinal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentID	Continuous	0	0 None		Never	Fixed	100	7119	0	0	0	0
Gender	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
Nationality	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
MaritalStatus	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
HomeEmirate(Emirates)_Code	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentTypeTypeCode	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
@1stAcademicPeriod	Ordinal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentDegreeDegreeCode	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentMajorMajor	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentMajorCIPMajorClassification	Ordinal	--	--		Never	Fixed	100	7119	0	0	0	0
StudentProgramProgramCode	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
ModeofStudyFTPT	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
EmploymentStatus	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
@12thGradeStreamStreamCode	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
OutgoingExchangestudentIndicator	Nominal	--	--		Never	Fixed	100	7119	0	0	0	0
Attrition	Flag	--	--		Never	Fixed	100	7119	0	0	0	0

4.2.3. Statistical Analysis

As discussed in Chapters 1 and 3, the selected RIT Dubai datasets are of Enrollments and Attrition from the academic year 2020-2021 until 2022-2023. The descriptive statistics below describes the different variables in the datasets and the relationships between various variables:

Student Data by Gender:

Male= 1708 (74%)

Female= 584 (25%)

Unknown= 11 (1%)

Percentage of Male > Percentage of Female Students

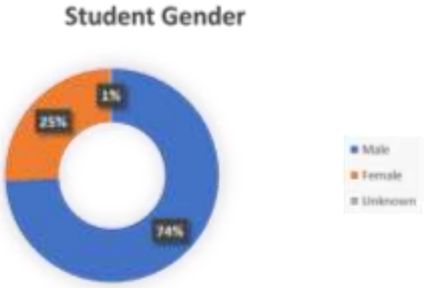


Figure 4- Student Data by Gender

Student Data by GPA:

Max GPA= 4.0

Min GPA= 0.0

Mean= 2.05

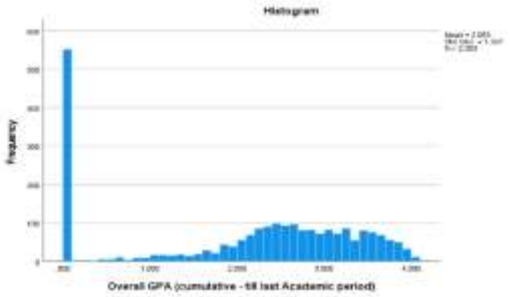


Figure 5- Student Data by GPA

Total credits registered by total credits completed:

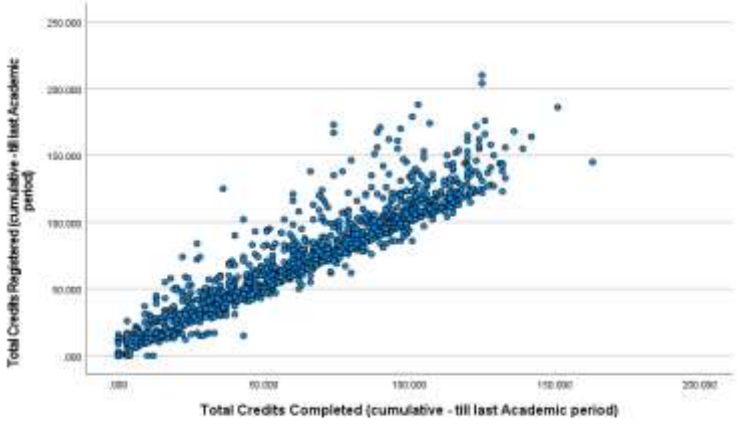


Figure 6-Credits Registered by Total Credits Completed

Overall GPA by Gender:

Female: 0-2.25

Male: 0-1.9

Unknown: 0-0.6

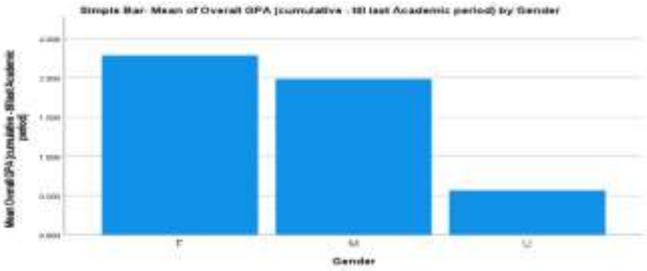
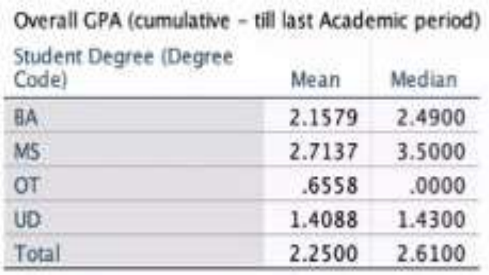


Figure 7- Overall GPA by Gender

Student Degree Level Compared to Overall GPA:



Overall GPA (cumulative – till last Academic period)		
Student Degree (Degree Code)	Mean	Median
BA	2.1579	2.4900
MS	2.7137	3.5000
OT	.6558	.0000
UD	1.4088	1.4300
Total	2.2500	2.6100

Figure 9-Student Degree Level Compared to Overall GPA

Students by Program Level:

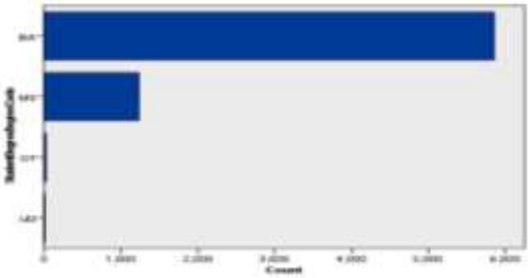


Figure 8-Students by Program Level

Student Levels:

Freshmen= 1113 (48%)

Junior= 383 (17%)

Sophomore= 438 (19%)

Senior=369 (16%)

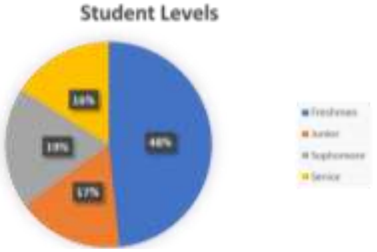


Figure 10-Student Levels

Students Type:

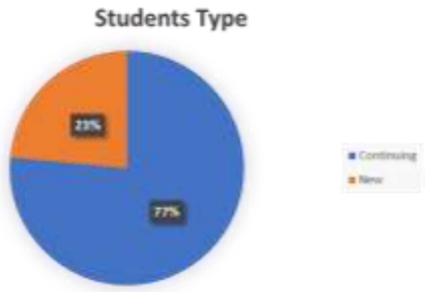


Figure 11-Students Type

Continuing= 1762 (77%)

New= 541 (23%)

Number of BS Students by Programs:

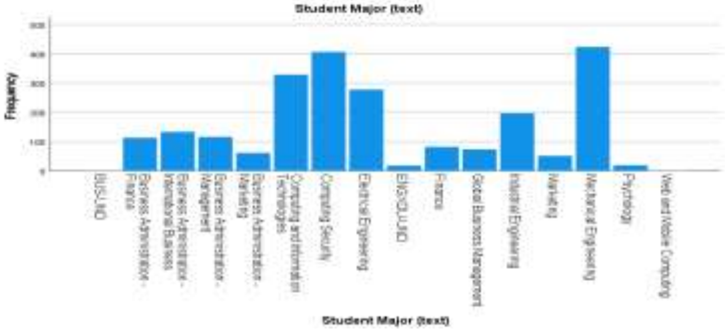


Figure 12-Number of BS Students by Programs

Number of Students by Mode of Study:

Full Time: 1902 (83%)

Part Time: 401 (17%)

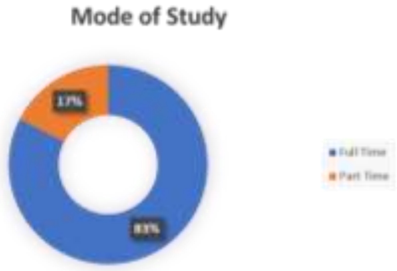


Figure 13-Number of Students by Mode of Study

Number of students by curriculum:

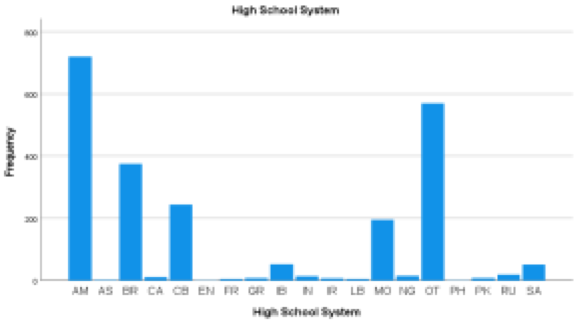


Figure 15-Number of Students by Curriculum

High School Exit Score:

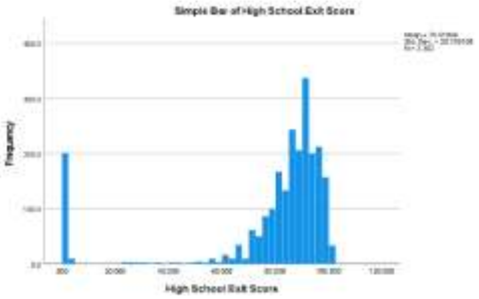
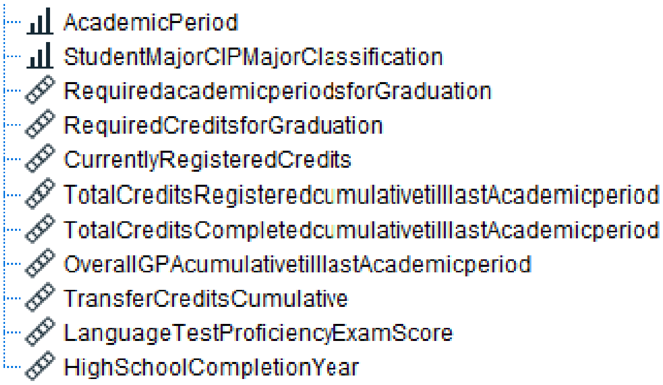


Figure 14-High School Exit Score

4.2.4. Dimensionality Reduction

For dimensionality reduction, Principal Components Analysis (PCA) technique was applied to reduce the number of variables and their impact on the dataset. PCA is a linear algebra technique for continuous attributes that finds new attributes (principal components) that are linear combinations of the original attributes, are orthogonal (perpendicular) to each other, and capture the maximum amount of variation in the data (Tan, Steinbach, & Kumar, 2014).

In this research, the 11 original attributes used for the PCA are:

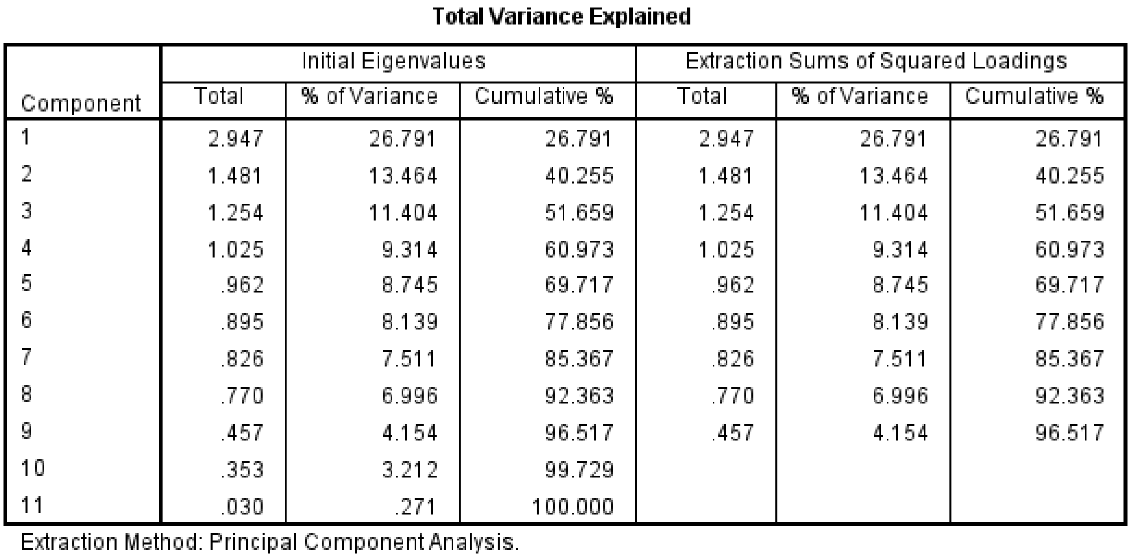


- AcademicPeriod
- StudentMajorCIPMajorClassification
- RequiredacademicperiodsforGraduation
- RequiredCreditsforGraduation
- CurrentlyRegisteredCredits
- TotalCreditsRegisteredcumulativetilllastAcademicperiod
- TotalCreditsCompletedcumulativetilllastAcademicperiod
- OverallGPAcumulativetilllastAcademicperiod
- TransferCreditsCumulative
- LanguageTestProficiencyExamScore
- HighSchoolCompletionYear

Figure 16- Attributes used for PCA

The table below shows the total variance to help choose the number of components defined by PCA.

Table 8-Total Variance (PCA)



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.947	26.791	26.791	2.947	26.791	26.791
2	1.481	13.464	40.255	1.481	13.464	40.255
3	1.254	11.404	51.659	1.254	11.404	51.659
4	1.025	9.314	60.973	1.025	9.314	60.973
5	.962	8.745	69.717	.962	8.745	69.717
6	.895	8.139	77.856	.895	8.139	77.856
7	.826	7.511	85.367	.826	7.511	85.367
8	.770	6.996	92.363	.770	6.996	92.363
9	.457	4.154	96.517	.457	4.154	96.517
10	.353	3.212	99.729			
11	.030	.271	100.000			

Extraction Method: Principal Component Analysis.

As shown in Table 8, with 1 component the variance was at 26.7%. With 6 components the variance increased to 77.8% and with 9 components the variance was at 96.5%. Hence, 9 components were chosen for dimension reduction.

The equations for each of the 9 PCA factors chosen are linear combinations of the 11 original attributes provided to the PCA:

Table 9- Equations for PCA Factors

Factor	Equation
Equation for Factor-1	-0.001279 * AcademicPeriod - 0.002948 * StudentMajorCIPMajorClassification + 0.009293 * RequiredacademicperiodsforGraduation + 0.02743 * RequiredCreditsforGraduation -0.0284 * CurrentlyRegisteredCredits + 0.007193 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.008347 * TotalCreditsCompletedcumulativetilllastAcademicperiod + 0.1662 * OverallGPAcumulativetilllastAcademicperiod + 0.01224 * TransferCreditsCumulative -0.0001927 * LanguageTestProficiencyExamScore + 0.0000004795 * HighSchoolCompletionYear + 254.2
Equation for Factor-2	0.0001033 * AcademicPeriod + 0.02891 * StudentMajorCIPMajorClassification -0.03309 * RequiredacademicperiodsforGraduation -0.1609 * RequiredCreditsforGraduation + 0.02267 * CurrentlyRegisteredCredits + 0.002432 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.002632 * TotalCreditsCompletedcumulativetilllastAcademicperiod + 0.0949 * OverallGPAcumulativetilllastAcademicperiod + 0.003747 * TransferCreditsCumulative -0.0001172 * LanguageTestProficiencyExamScore -0.00000004179 * HighSchoolCompletionYear -1.658
Equation for Factor-3	0.005825 * AcademicPeriod -0.006201 * StudentMajorCIPMajorClassification + 0.005965 * RequiredacademicperiodsforGraduation + 0.02542 * RequiredCreditsforGraduation + 0.09552 * CurrentlyRegisteredCredits + 0.003053 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.004105 * TotalCreditsCompletedcumulativetilllastAcademicperiod + 0.2894 * OverallGPAcumulativetilllastAcademicperiod + 0.008207 * TransferCreditsCumulative + 0.0002395 * LanguageTestProficiencyExamScore -0.000001104 * HighSchoolCompletionYear -1182.2
Equation for Factor-4	-0.0006612 * AcademicPeriod + 0.01709 * StudentMajorCIPMajorClassification + 0.2176 * RequiredacademicperiodsforGraduation + 0.02065 * RequiredCreditsforGraduation + 0.0006228 * CurrentlyRegisteredCredits -0.0004149 * TotalCreditsRegisteredcumulativetilllastAcademicperiod -0.000565 * TotalCreditsCompletedcumulativetilllastAcademicperiod + 0.001602 * OverallGPAcumulativetilllastAcademicperiod -0.001813 * TransferCreditsCumulative + 0.0000183 * LanguageTestProficiencyExamScore -0.00000026 * HighSchoolCompletionYear + 128.8
Equation for Factor-5	0.001477 * AcademicPeriod + 0.001625 * StudentMajorCIPMajorClassification + 0.0004923 * RequiredacademicperiodsforGraduation -0.004303 * RequiredCreditsforGraduation -0.03343 * CurrentlyRegisteredCredits -0.000777 * TotalCreditsRegisteredcumulativetilllastAcademicperiod -0.0006545 * TotalCreditsCompletedcumulativetilllastAcademicperiod -0.2082 * OverallGPAcumulativetilllastAcademicperiod + 0.1148 * TransferCreditsCumulative + 0.00006977 *LanguageTestProficiencyExamScore -0.0000002929 * HighSchoolCompletionYear -297.1
Equation for Factor-6	0.001498 * AcademicPeriod + 0.005768 * StudentMajorCIPMajorClassification -0.004953 * RequiredacademicperiodsforGraduation -0.01571 * RequiredCreditsforGraduation -0.09545 *

	CurrentlyRegisteredCredits + 0.003408 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.00345 * TotalCreditsCompletedcumulativetilllastAcademicperiod -0.006749 * OverallGPAcumulativetilllastAcademicperiod -0.01325 * TransferCreditsCumulative + 0.00135 * LanguageTestProficiencyExamScore + 0.0000003324 * HighSchoolCompletionYear -300.5
Equation for Factor-7	0.008913 * AcademicPeriod + 0.001153 * StudentMajorCIPMajorClassification + 0.0249 * RequiredacademicperiodsforGraduation -0.00667 * RequiredCreditsforGraduation + 0.02989 * CurrentlyRegisteredCredits + 0.003216 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.002138 * TotalCreditsCompletedcumulativetilllastAcademicperiod -0.3346 * OverallGPAcumulativetilllastAcademicperiod -0.01357 * TransferCreditsCumulative -0.0002538 * LanguageTestProficiencyExamScore + 0.000001725 * HighSchoolCompletionYear -1801.4
Equation for Factor-8	-0.005451 * AcademicPeriod -0.0003948 * StudentMajorCIPMajorClassification + 0.003687 * RequiredacademicperiodsforGraduation + 0.01761 * RequiredCreditsforGraduation + 0.1858 * CurrentlyRegisteredCredits -0.00004545 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.0004115 * TotalCreditsCompletedcumulativetilllastAcademicperiod -0.03632 * OverallGPAcumulativetilllastAcademicperiod + 0.02442 * TransferCreditsCumulative + 0.0005776 * LanguageTestProficiencyExamScore + 0.000001346 * HighSchoolCompletionYear + 1096.1
Equation for Factor-9	-0.00003673 * AcademicPeriod + 0.04916 * StudentMajorCIPMajorClassification -0.09695 * RequiredacademicperiodsforGraduation + 0.2929 * RequiredCreditsforGraduation + 0.01803 * CurrentlyRegisteredCredits + 0.003279 * TotalCreditsRegisteredcumulativetilllastAcademicperiod + 0.002366 * TotalCreditsCompletedcumulativetilllastAcademicperiod -0.2027 * OverallGPAcumulativetilllastAcademicperiod -0.01016 * TransferCreditsCumulative -0.00006647 * LanguageTestProficiencyExamScore -0.0000005317 * HighSchoolCompletionYear -29.69

4.2.5. Visualization of Key Features

GPA of Students Leaving: Approximately 0 - 2.25

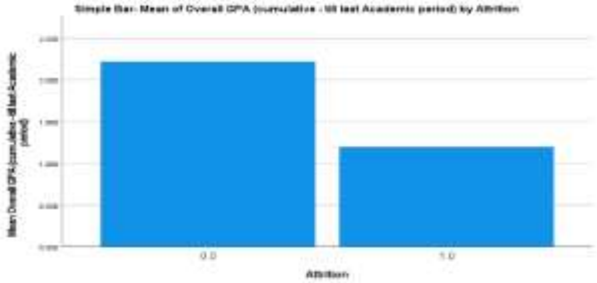
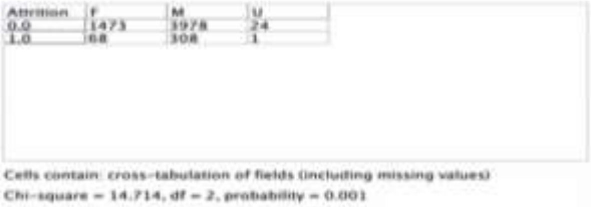


Figure 17-GPA of Students Leaving

Student Attrition by Gender:

The number of Male attritors is more than Female. The P value of the chi-square is 0.001 which shows a significant relationship between Gender and Attrition.



Attrition	F	M	U
0.0	1473	3978	24
1.0	68	308	1

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 14.714, df = 2, probability = 0.001

Figure 18- Student Attrition by Gender

Student Attrition by Major:

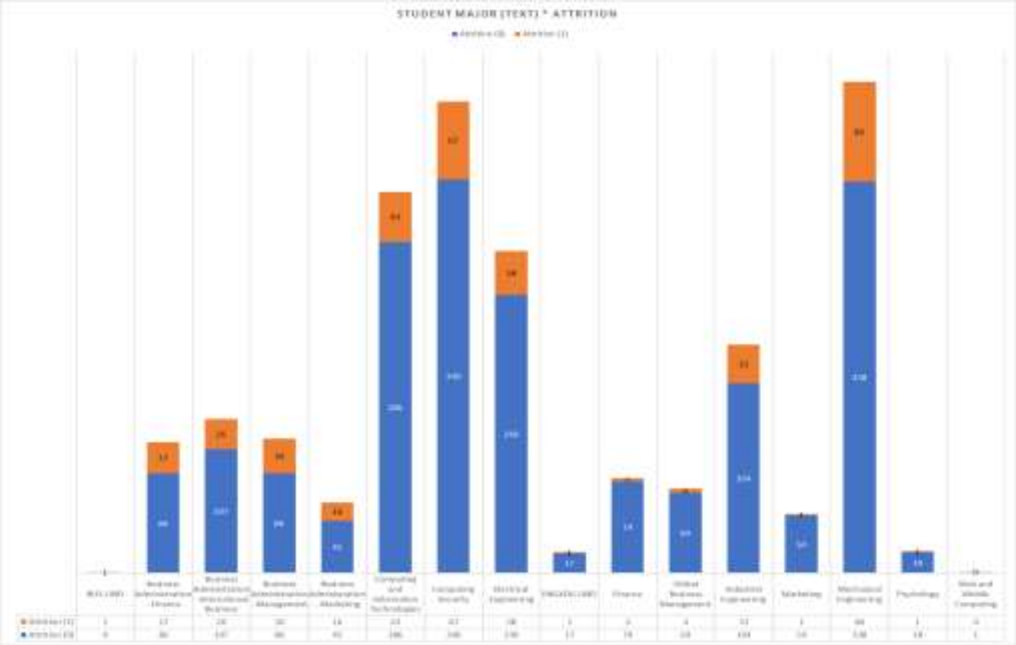


Figure 19- Student Attrition by Major

The total number of attritors were 377. BS in Mechanical Engineering has the **highest number of attritors (86) at 22.8%** followed by Computing program- Computing Security with 17.8% (67) and Computing and Information Technologies with 11.4% (43). Also, Psychology, Marketing and Finance majors have the **least number of attritors**.

Boxplot to show the Attrition by High School Exit Score:

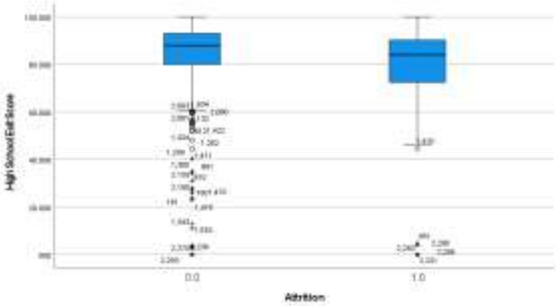


Figure 20- Attrition by High School Exit Score

The boxplot in Figure 20 shows attrition by High School Exit Score. It shows the 25th percentile, 75th percentile, mean, and outliers.

Boxplot to show the Attrition by Total Credits Completed (cumulative- till last academic period

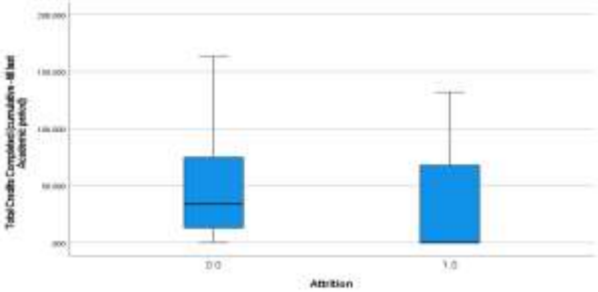


Figure 21-Attrition by Total Credits Completed till last academic period.

As shown in the Boxplot, the Total Credits Completed has an impact on Attrition. It is worth noting that there are no outliers.

Student Level (Level Code) by Attrition:

Table 10- Student Level Code by Attrition

		Attrition		Total
		0.0	1.0	
Student Level(Level Code)	FR	889	224	1113
	JR	347	36	383

	SP	399	39	438
	SR	291	78	369
Total		1926	377	2303

The chi-square is <.001 which shows the Student Level (Level Code) is a significant factor.

Table 11-Chi-Square: Student Level (Level Code)

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	49.022[a]	3	<.001
Likelihood Ratio	52.904	3	<.001
N of Valid Cases	2303		

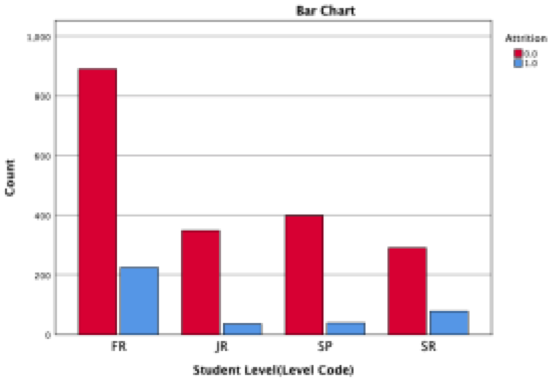


Figure 22- Bar Chart: Student Level (Level Code) by Attrition

Program (Program Code) by Attrition:

Table 12-Program Code by Attrition

		Attrition		Total
		0.0	1.0	
Program (Program Code)		1926	1	1927
	59.BA.1210	0	43	43

	59.BA.1211	0	66	66
	59.BA.1212	0	37	37
	59.BA.1213	0	29	29
	59.BA.1214	0	32	32
	59.BA.1215	0	28	28
	59.BA.1216	0	35	35
	59.BA.1217	0	17	17
	59.BA.1218	0	86	86
	59.NA.9999	0	3	3
Total		1926	377	2303

The chi-square is <.001 which shows the Program Code is a significant factor.

Table 13-Chi-Square: Program Code

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2295.699[a]	10	<.001
Likelihood Ratio	2036.016	10	<.001
N of Valid Cases	2303		

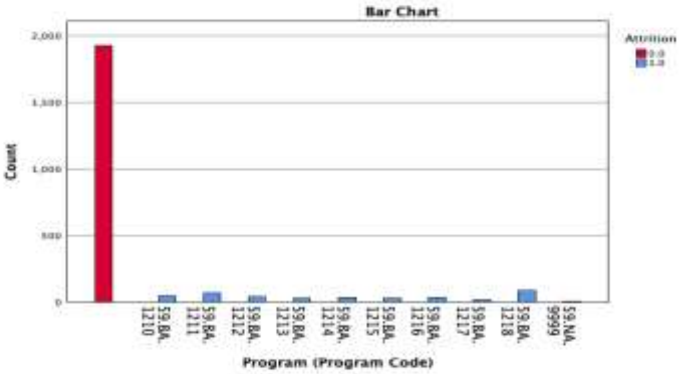


Figure 23- Bar Chart: Program Code by Attrition

Student Major (text) by Attrition:

Table 14- Student Major by Attrition

		Attrition		Total
		0.0	1.0	
Student Major (text)	BUS-UND	0	1	1
	Business Administration - Finance	86	27	113
	Business Administration - International Business	107	26	133
	Business Administration - Management	86	30	116
	Business Administration - Marketing	45	16	61
	Computing and Information Technologies	286	43	329
	Computing Security	340	67	407
	Electrical Engineering	240	38	278
	ENGXDU-UND	17	1	18
	Finance	79	3	82
	Global Business Management	69	4	73
	Industrial Engineering	164	33	197
	Marketing	50	1	51
	Mechanical Engineering	338	86	424
	Psychology	18	1	19
	Web and Mobile Computing	1	0	1
Total		1926	377	2303

The chi-square is <.001 which shows the Student Major (text) is a significant factor.

Table 15- Chi-Square: Student Major

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	58.776[a]	15	<.001
Likelihood Ratio	65.806	15	<.001
N of Valid Cases	2303		

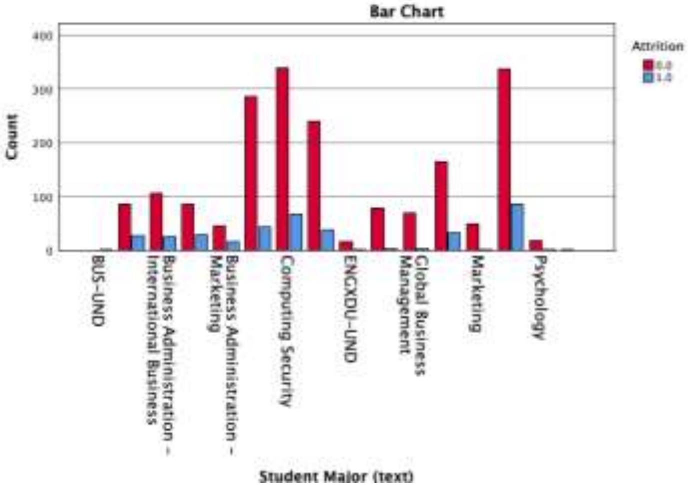


Figure 24- Bar Chart: Student Major by Attrition

Hypothesis Test for Factors Containing Numerical Values:

To evaluate the importance of the numerical values and the attrition the non-parametric Mann-Whitney U Test was used. It is a non-parametric version of the t-test that can be used when the aim is to show a difference between two groups in the value of an ordinal, interval, or ratio variable (McIntosh, Sharpe, & Lawrie, 2010). The Mann–Whitney test is used to compare two independent groups. It can detect differences in the spread as well as the location (median) of two variables, even when we have similar medians. So, when presenting the results of Mann–Whitney tests, the median of each group should be presented along with a description of the skewness of each sample for example with a box plot.

Table 16- Hypothesis Test Summary

Hypothesis Test Summary			
Null Hypothesis	Test	Sig.^{a,b}	Decision
The distribution of Language Test Proficiency Exam Score is the same across categories of Attrition.	Independent-Samples Mann-Whitney U Test	.467	Retain the null hypothesis.
The distribution of Factor2 is the same across categories of Attrition.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.
The distribution of Factor4 is the same across categories of Attrition.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

As shown in table 16 above, the distribution of Factor2 and Factor4 is different across categories of Attrition. However, for the distribution of Language Test Proficiency Exam the null hypothesis has been retained which means that the distribution is the same across categories of Attrition. Hence, we will only show the distribution of Factor2 and Factor4.

Factor2 by Attrition:

Table 17-Factor2 by Attrition

Independent-Samples Mann-Whitney U Test Summary	
Total N	2303
Mann-Whitney U	298698.000
Wilcoxon W	369951.000
Test Statistic	298698.000
Standard Error	11807.268
Standardized Test Statistic	-5.450
Asymptotic Sig. (2-sided test)	<.001

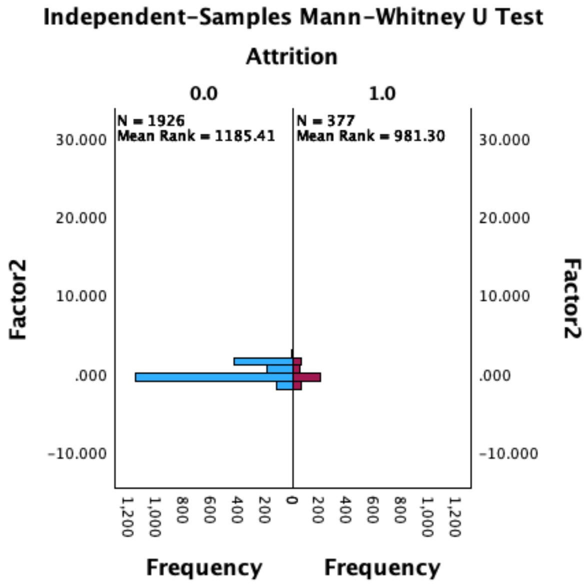


Figure 25-Frequency for Factor2

Factor4 by Attrition:

Table 18- Factor4 by Attrition

Independent-Samples Mann-Whitney U Test Summary	
Total N	2303
Mann-Whitney U	408913.000
Wilcoxon W	480166.000
Test Statistic	408913.000
Standard Error	11807.268
Standardized Test Statistic	3.884
Asymptotic Sig. (2-sided test)	<.001

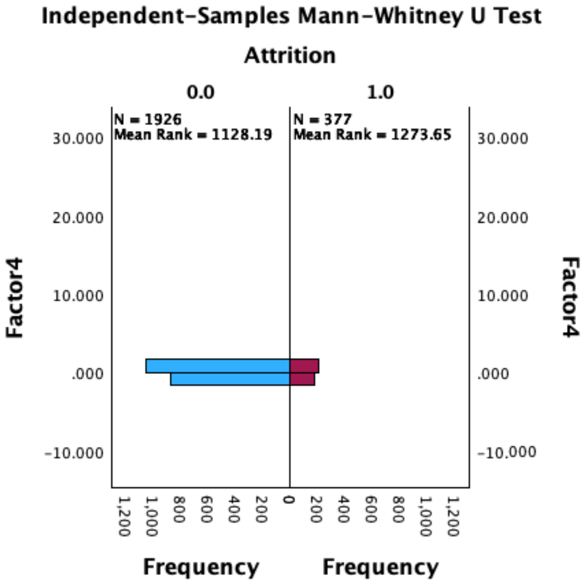


Figure 26-Frequency for Factor4

4.3 Machine Learning Model Development

4.3.1. Detailed Explanation of the Chosen Input

Two techniques were tested for feature selection. The feature selection node of the SPSS modeler was used to identify the fields that are most important for the analysis. For example, we were trying to predict student attrition based on several factors. Which factors were the most likely to be important?

Feature Selection Technique #1 consisted of three steps:

- **Screening:** Removes unimportant and problematic inputs and records, or cases such as input fields with too many missing values or with too much or too little variation to be useful.
- **Ranking:** Sorts remaining inputs and assigns ranks based on importance.
- **Selecting:** Identifies the subset of features to use in subsequent models—for example, by preserving only the most important inputs and filtering or excluding all others.

Table 19 shows the important features selected through feature selection technique #1.

Table 19- Feature Importance- Technique #1

Field	Measurement	Importance	Value
StudentTypeCode	Nominal	Important	1.0
$F-Factor-3	Continuous	Important	1.0
OverallGPAcumulativetilllastAcademicperiod	Continuous	Important	1.0
AcademicPeriod	Ordinal	Important	1.0
HighSchoolCountryCode	Nominal	Important	1.0
StudentLevelCode	Nominal	Important	1.0
ModeofStudyFTPT	Nominal	Important	1.0
1stAcademicPeriod	Ordinal	Important	1.0
$F-Factor-5	Continuous	Important	1.0
Nationality	Nominal	Important	1.0
CurrentlyRegistredCredits	Continuous	Important	1.0
TotalCreditsCompletedcumulativetilllastAcademicperiod	Continuous	Important	1.0
HighSchoolEquivalencyNew	Nominal	Important	1.0
TotalCreditRegistredcumulativetilllastAcademicperiod	Continuous	Important	1.0
HighSchoolSystem	Nominal	Important	1.0
$F-Factor-2	Continuous	Important	1.0
Gender	Nominal	Important	1.0
StudentMajortext	Nominal	Important	1.0
$F-Factor-1	Continuous	Important	0.998
TransferCreditsCumulative	Continuous	Important	0.995
$F-Factor-7	Continuous	Important	0.994

The second feature selection technique used was Random Forest for selecting important factors using 8 variables as input with Attrition as the Target. Figure 2 shows the important features selected through this technique.

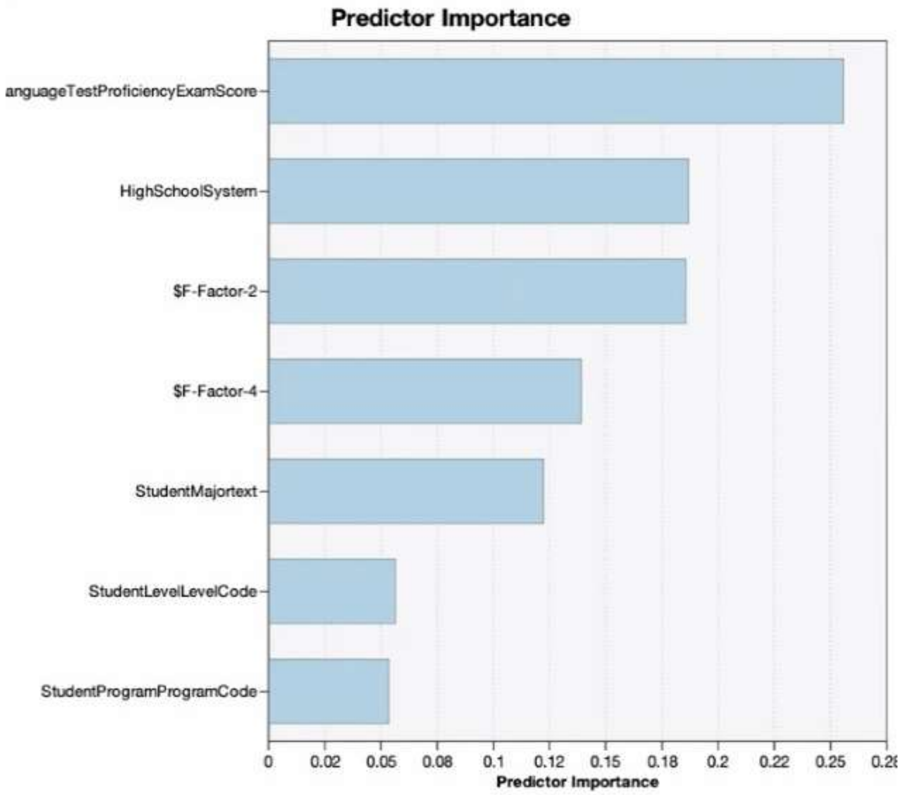


Figure 27-Feature Importance- Technique #2

After using the two automatic feature selection techniques, we have LanguageTestProficiencyExamScore, StudentProgramCode, and $F- Factor 4 which are unique to feature selection technique #2.

4.3.2. Detailed Explanation of the Chosen ML Models

Machine Learning Algorithms used in the research are provided below:

i. **Random Trees**

Random Trees ML algorithm is an effective classifier created by a collection of tree predictors in such a way that each tree is reliant on independent values of a sampled random vector (Breiman, 2001).

ii. **Logistic Regression**

According to (Giuseppe, 2018), Logistic Regression is used to estimate the relationship between a categorical dependent variable and one or more independent variables. It is used to estimate the likelihood that an event will occur.

iii. **Linear Support Vector Machines (LSVM)**

Linear Support vector machine (LSVM), one of the four methods of SVM, is a very fast, simple algorithm used to handle problems with very large dimensional input spaces (Lu, S., & Wang, X. 2004).

iv. **Neural Networks**

Neural network aims to replicate the structure and operation of the human brain. As described by (Giuseppe, 2018), Artificial Neural Networks (ANN) or Neural network is a directed or recurrent computational structure that connects an input layer to an output layer.

4.3.3. Validation and Testing Procedures

4.3.3.1. Data Partitioning

In order to perform the prediction, the data sets were partitioned into two- 70% for training corresponding to 3785 students and 30% for testing corresponding to 1690 students.

4.3.3.2. Data Balancing

The dataset was imbalanced- 93.56% values were 0's and only 6.44% were 1's, so balancing techniques were used.

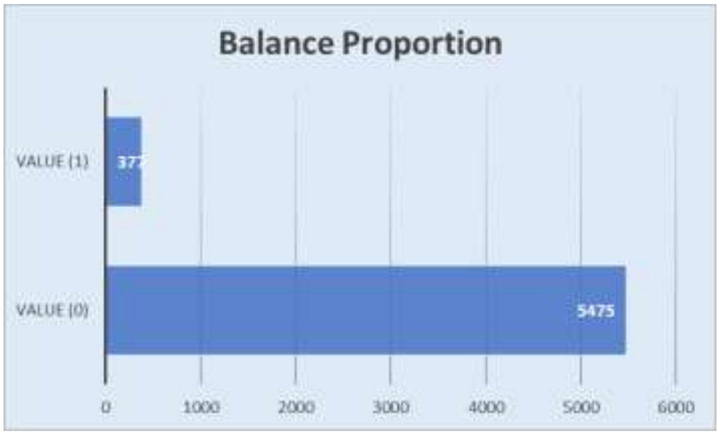


Figure 28-Data Balancing

The training dataset was balanced to have the same number of 1's and same number of 0's as shown in the figure above. Undersampling technique was used instead of oversampling.

Table 20- Balance Directives

Balance Directives	
Factor	**Condition**
0.0688	Attrition= 0
1	Attrition= 1

As shown in Table 20, 0 was multiplied with 0.688 to reduce this percentage. Similarly, we performed the same technique for the other factors to achieve the following results:

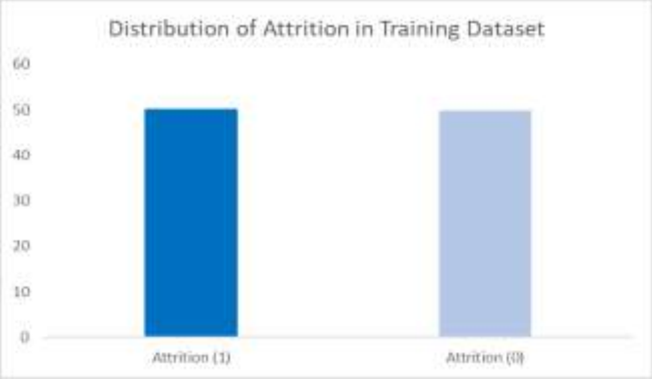


Figure 29- Data Partitioning

4.3.3.3. Evaluation Metrics Used to Assess Model Performance

Predicting student attrition is a binary classification problem in ML. To evaluate the results of this research study, classification metrics and performance metrics such as accuracy (ACC), AUC (Area under receiver operating characteristics curve (AUC), Precision (PR and Recall (Rec) were applied (Powers, 2020). The accuracy, sensitivity, and specificity are based on the confusion matrix (Chung & Lee, 2019). In a confusion matrix, a 2 by 2 cross-table is used to show the performance of the models for the binary classification. The class of interest or Attrition is labelled as positive, and the other class is labelled as negative. As shown in table 21 below, the actual positive and negative are on the column, and the positive and negative that are predicted are on the row:

Table 21- Confusion Matrix

Confusion Matrix			
		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

According to (Chung & Lee, 2019), in the confusion matrix, the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) are the cases when the actual and predicted classes are both positive, the cases when the actual and predicted classes are both negative, the cases when the actual class is negative, but predicted class is positive, and the cases when the actual class is positive, but the predicted class is negative. The accuracy is defined as the proportion of correct predictions over the total number of predictions made by the model, and it is represented in the confusion matrix as (TP + TN)/ (TP + FP + FN + TN). The sensitivity (true positive rate) is the proportion of those predicted as positive (Student Attritors) among the true positive, and is defined as TP/ (TP + FN) in the confusion matrix. The specificity (true negative rate) is the proportion of those predicted as negative (non-attritors) among the true negative, and is defined as TN/ (TN + FP) in the confusion matrix.

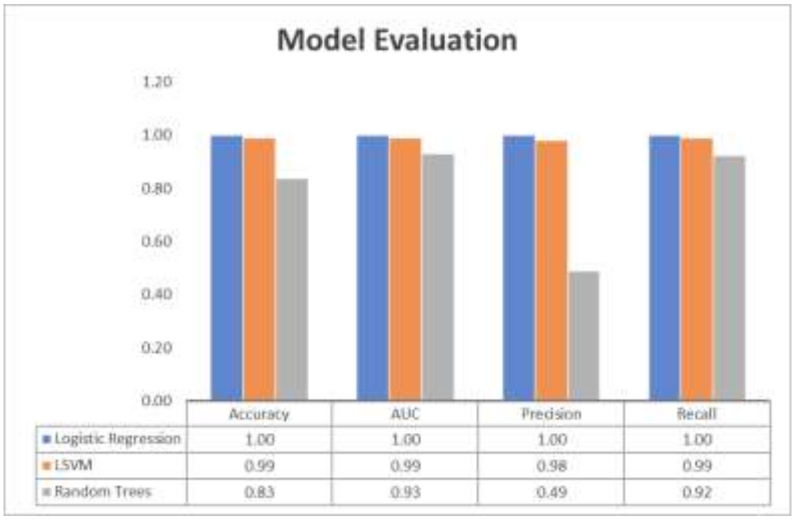
Accuracy is a commonly used performance indicator to evaluate the results of the models. However, we cannot always rely only on accuracy because it can lead to misinterpretation as the model can predict only the dominant class and ignore the minor class (Riquelme, Lücken, & Baran, 2015). The Receiver Operating Characteristics (ROC) curve, which can cover the entirety of a classification method's prediction performance for all classification thresholds, is represented by the AUC, a performance metric that is widely used for machine learning classification models (Muschelli, 2020). It indicates the extent to which the model can distinguish between classes. The accuracy of the ML model was measured using AUC. According to (Koizumi, Murata, Harada, Saito, & Uematsu, 2019), AUC is known as a popular metric used for evaluating the effectiveness of binary classification models, capturing the trade-off between sensitivity (TPR- True Positive Rate) and specificity (FPR- False Positive Rate) among various thresholds. When the AUC is higher the model has a better ability to predict 0s as 0s and 1s as 1s. AUC has a range from 0.5-

1.0 with a value of 0.5 denoting a random classification model and a value of 1.0 denoting a perfect classification model. When the value of AUC is near 1.0, it represents a better performance of the model while a value less than 0.5 represents poor performance and the steepness of ROC should be high as it represents TPR with less FPR (Koizumi, Murata, Harada, Saito, & Uematsu, 2019).

4.3.4. Results

4.3.4.1. Presentation of the Experimental Results

8 variables or features were automatically selected through Random Forest and Logistic Regression, Random Trees and LSVM models were applied to these features. The figure below shows the performance and accuracy for all three models.



	Accuracy	AUC	Precision	Recall
■ Logistic Regression	1.00	1.00	1.00	1.00
■ LSVM	0.99	0.99	0.98	0.99
■ Random Trees	0.83	0.93	0.49	0.92

Figure 30- Model Evaluation

4.3.4.2. Comparison of different machine learning models

Logistic Regression proved to be the best model with accuracy, AUC, precision and recall at 1.0. However, it is noteworthy that LSVM model also had almost similar accuracy, AUC, precision and recall at approximately 0.98 – 0.99. However, Logistic Regression was chosen to be a better model as the Precision was at 1.0. The models were tested using 34 variables and 8 variables and the final results were improved with 8 variables.

4.3.4.3. Evaluation of Predictor Importance

The Predictor importance can be found by computing the reduction in the variance of the target (attrition) assigned to each predictor through the sensitivity analysis (Saltelli, Tarantola, F., Campolongo, & Ratto, 2004).

The following notation was used in this study:

Table 22- Predictor Importance Notations

y	Target
X_j	Predictor, where j=1,….k
k	The number of predictors
$Y=f(X_1, X_2,\dots\dots X_j)$	Model for Y based on predictors X_1 through X_k

Predictors are ranked based on the sensitivity measure as shown below:

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i))}{V(Y)}$$

where $V(Y)$ is the unconditional output variance. In the numerator, the expectation operator E calls for an integral over; that is, overall factors but, then the variance operator V implies a further integral over. Predictor importance is then computed as the normalized sensitivity.

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j}$$

where S_i is the appropriate measure of sensitivity to rank the predictors in the order of importance. The importance measure S_i is the first-order sensitivity measure, which is accurate if the set of the input factors (X_1, X_2,…, X_k) is orthogonal/independent (a property of the factors), and the model is additive; i.e., the model does not include interactions (a property of the model) between the input factors (Saltelli, Tarantola, F., Campolongo, & Ratto, 2004). For any combination of interaction and non-orthogonality among factors, S_i is still the proper measure of sensitivity to rank the input factors in order of importance, but there can be inaccuracy due to the presence of interactions or/and non-orthogonality. For a better estimation of S_i, the size of the dataset should be a few hundred at least, or else S_i can be biased and the improvement of the importance measure can be done through bootstrapping (Saltelli, Tarantola, F., Campolongo, & Ratto, 2004).

4.3.4.4. Predictor Importance of the Best Model

The graphs below show the most important factors selected by LSVM and Random Trees for student attrition based on their importance:

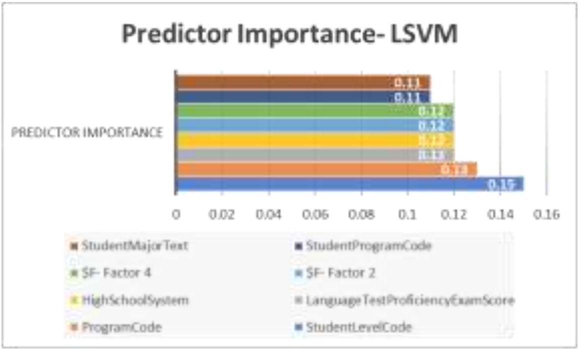


Figure 32- Predictor Importance- LSVM

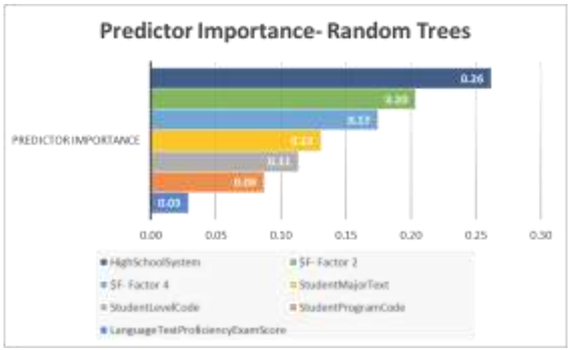


Figure 31-Predictor Importance- Random Trees

For example, Figure 31 shows that for LSVM, the most important factor that contributes highly to student attrition is the StudentLevel Code with a sensitivity of 0.15. As shown in Figure 32, for Random Trees, the most important factor that contributes highly to student attrition is HighSchool System with a sensitivity of 0.26.

4.3.4.5. Analysis of ROC curves and AUC values

The Figure below shows the ROC curve for the Logistic Regression model. As discussed in section 4.3.3.3, the steepness of the ROC curve is high, representing high TPR and low FPR (Koizumi, Murata, Harada, Saito, & Uematsu, 2019).

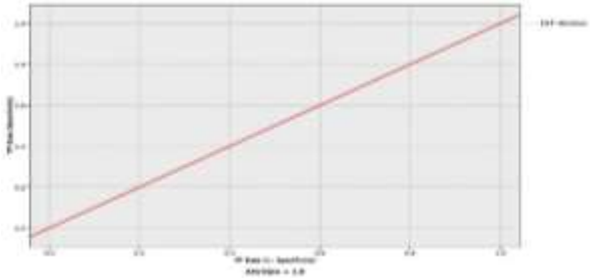


Figure 33- ROC Curve for Logistic Regression

Chapter 5 Discussion

5.1 Research Findings

As discussed in Section 1, this research aims to explore the use of machine learning algorithms such as Random Forest, LSVM, Random Trees, Logistic Regression and Neural Networks to develop an AI model for predicting student attrition in higher education institutions (HEI).

The primary research question aimed at the performance of ML algorithms performance in predicting student attrition in HEIs. As shown in section 4.3.4, Logistic Regression was the most successful model with consistent performance, accuracy and AUC.

The second research question aimed at the techniques that can used to enhance the data quality of the attrition and enrollment datasets. As discussed in section 4.2.2, missing values in the variables were imputed to improve the data quality. Further, in section 4.2.4, to achieve the dimensionality reduction PCA technique was applied to reduce the number of variables and their impact on the datasets. 9 components were chosen for dimension reduction as the total variance was achieved at 96.5%. Consequently, this helped to achieve the best accuracy through Logistic Regression model.

The third research question aimed at applying feature selection techniques to assist in automating the identification of the significant predictors of student attrition at RIT Dubai. As discussed in section 4.3.1, in this study, two feature selection techniques were explored. The first technique was based on Screening, ranking and selection and the second technique was through Random Forests. Both techniques facilitated finding the most important factors/ features for student attrition based on their importance. Similar to the research conducted by (Ahmad Tarmizi, Mutalib, Abdul Hamid, & Abdul Rahman, 2019), the variables such as gender, program code (defining the student major), high school grade and GPA were considered to be the most effective on the student's attrition decision. In this research, further significant factors such as Student Major, Language Test Proficiency Score (English Score- IELTS/ TOEFL), High School Equivalency (defining the authentication of their HS credentials), Student Type Code (describing if the student was continuing or new), Academic Period, Student Level (describing level of the student- Freshmen, Junior, Senior, Sophomore, Master), Mode of study (Full-time/ Part-time), Nationality, Currently Registered Credits, High School System, Total Credits Completed Cumulative, Total Credits Registered Cumulative and Transfer Credits Cumulative were also found to be very effective in predicting student attrition. These important features were applied on all the explored models. The most significant result in this research was achieved through the outstanding performance of the Logistic Regression model in predicting student attrition by achieving an accuracy of 100%. Moreover, this model also offered the same performance through AUC, precision, and recall at 1.0. This indicates the effectiveness of the Logistic Regression model in the prediction of student attrition.

The purpose of this research was to highlight the efficiency and competence of ML in predicting student attrition. (Jadric, Garaca, & Cukusic, 2010) applied Data Mining methods such as logistic regression, decision trees, and neural networks in which decision trees demonstrated greater capability to predict student attrition. Additionally, (Chung & Lee, 2019) applied random forests ML model to predict students' dropouts and the model showed exceptional performance with 95% accuracy and AUC at 97%. However, the study was limited to the research of descriptive features and potential inaccuracy in the weights used in calculating the features of the model. Similarly, (Kemper, 2020) applied logistic regression and decision trees for predicting student dropout and both methods provided high prediction accuracies of more than 83%. (Delen, 2010) explored the classification methods such as artificial neural networks, decision trees, support vector machines and logistic regression and ensemble techniques such as bagging, boosting and information fusion to predict and explain the reasons behind freshmen student attrition. Ensembles provided better results and the balanced dataset provided better prediction results. However, this research was limited to institutional data pertaining to freshman students. Consistent with the literature, Logistic Regression has proved to be widely used by researchers in the field of educational data mining including predicting student attrition. Through this research, we can clearly distinguish the Logistic Regression model outperforming all other ML models by offering superior results in predicting student attrition. Moreover, unlike previous research with lower sample sizes (which focused on a specific discipline, year, or freshman students data), this research focused on enrolment and attrition data for all the Bachelor level students from three academic years at RIT Dubai. This enabled greater accuracy in obtaining valid results through the Logistic Regression ML model.

In conclusion, the findings of this research study aligns with the broader literature on predictive modeling for student attrition. This research contributes to the body of knowledge aimed at addressing student attrition and, consequently, improving student retention efforts in HEIs by showcasing the efficiency of Logistic Regression, highlighting the need for data quality improvement, and utilizing feature selection techniques.

Chapter 6 Conclusion

6.1 Conclusion

Predicting student attrition is important for the students as well as the higher education institutions. This research paper presents an analysis of different machine learning techniques applied to predict student attrition within universities or higher education institutions. The analysis was conducted on three academic years data that was available at the time of enrolment and attrition data for all the students that are at the Bachelor degree level at RIT Dubai. The accomplished analysis of the model's performance takes into account the composition of the two unbalanced datasets. Considering predictions with the available enrolment and attrition datasets made the task difficult due to multiple missing values. Despite all the problems encountered, this research makes it possible to predict student attrition to help the institution to improve the student’s academic performance at early stages of their journey and also enables them to continuously monitor the students during their academic career.

With the help of various imputation techniques and PCA the data quality was enhanced. Additionally, two feature selection techniques were used in this research which assisted in automating the identification of the significant factors of student attrition at the institution. The first technique provided with 24 features that did not provide good results. For example, the Logistic regression model provided a precision of 0.16. However, when the features were selected using Random Forest, 8 best variables were selected. This enabled to improve the final performance of the models. As discussed in section 4.8 and 5.1 above, Logistic Regression model provided 1.0 accuracy, AUC, precision and recall. On the basis of these available predictions, the institution can decide on how and when to act to support the students. The final results and findings positively indicate that prediction of student attrition is possible to enable prompt mitigation of this issue.

6.2 Recommendations

The results of this research will be shared with the RIT Dubai management to be used for predicting student attrition and increasing student retention. The recommendation would be to use these predictions to improve the students’ performance by providing specific support such as providing additional academic support courses, tutoring sessions, personalized academic plans etc. Through this research, I strongly believe that this is an effective solution to decrease student attrition and ultimately increase student retention in higher education institutions in the UAE. However, it should be noted that the predicted factors may change over the academic years for different students.

6.3 Future Work

Further integration with the Student Information System (SIS) to automatically update the datasets will need deploying the software which will need research funding. Constant monitoring is essential to evaluate the effectiveness of the model to ensure accuracy in the prediction of the factors affecting attrition. Another enhancement would be the inclusion of data for Graduate degree level students as this research is limited to students from Bachelor degree level.

References

1. Ahmad Tarmizi, S. S., Mutalib, S., Abdul Hamid, N. H., & Abdul Rahman, S. (2019). A Review on Student Attrition in Higher Education Using Big Data Analytics and Data Mining Techniques.

2. Bäckman, O., & Nilsson, A. (2016). Long-term consequences of being not in employment, education or training as a young adult. Stability and change in three Swedish birth cohorts. *European Societies, 18*(2), 136-157. doi:https://doi.org/10.1080/14616696.2016.1153699

3. Barefoot, B. O. (2007). Higher education's revolving door: confronting the problem of student dropout in US colleges and universities. *19*(1), 9-18. doi:https://doi-org.ezproxy.rit.edu/10.1080/0268051042000177818

4. Barramuño, M., Meza-Narváez, C., & Gálvez-García, G. (2022). Prediction of student attrition risk using machine learning. *Journal of Applied Research in Higher Education, 14*(3), 974-986. doi:DOI:10.1108/JARHE-02-2021-0073

5. Breiman, L. (2001). Random Forests. *Machine Learning, 45*, pages5–32. doi:https://doi.org/10.1023/A:1010933404324

6. Catterall, J. S. (1985). *On the social costs of dropping out of schools. (Report No. 86-SEPT-3).* CA: Stanford, CA: Stanford University, Center for Educational Research.

7. Chauvet, G., Deville, J., & & Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika, 98*(2), 459-471. doi:http://www.jstor.org/stable/23076163

8. Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review, 96*, 346-353. doi:https://doi.org/10.1016/j.childyouth.2018.11.030

9. Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student Dropout Prediction. *Artificial Intelligence in Education (AIED)*, 129-140. doi:https://doi-org.ezproxy.rit.edu/10.1007/978-3-030-52237-7_11

10. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *49*(4), 498-506. doi:https://doi.org/10.1016/j.dss.2010.06.003

11. Giuseppe, B. (2018). *Machine Learning Algorithms : Popular Algorithms for Data Science and Machine Learning* (2nd Edition ed.). Birmingham, UK: Packt Publishing, Limited. Retrieved from http://ebookcentral.proquest.com/lib/rit/detail.action?docID=5504925.

12. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge: MIT Press.

13. Graham, J. (2012). *Missing data: Analysis and design* (1 ed.). New York: Springer New York, NY. doi:https://doi.org/10.1080/14616696.2016.1153699

14. Helbling, L. A., & Sacchi, S. (2014). Scarring effects of early unemployment among young workers with vocational credentials in Switzerland. *Empirical Research in Vocational Education and Training, 6*(1), 12. doi:https://doi.org/10.1186/s40461-014-0012-2

15. Jadric, M., Garaca, Z., & Cukusic, M. (2010). Student Dropout Analysis with Application of Data Mining Methods. *Management: Journal of Contemporary Management Issues, 15*(1), 31-46.

16. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* New York: Springer New York, NY. doi:https://doi-org.ezproxy.rit.edu/10.1007/978-1-4614-7138-7

17. Kelleher, J., Namee, B. M., & D'arcy, A. (2020). Fundamentals of machine learning for predictive data analytics: Algorithms, Worked Examples, and Case Studies. In J. D. Kelleher, B. M. Namee, & A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data* (pp. 1-29). London, England: MIT Press.

18. Kemper, L. (2020). Predicting student dropout: A machine learning approach. *European journal of higher education, 10*(1), 28. doi:10.1080/21568235.2020.1718520

19. Koizumi, Y., Murata, S., Harada, N., Saito, S., & Uematsu, H. (2019). SNIPER: Few-shot learning for anomaly detection to minimize false-negative rate with ensured true-positive rate. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 915-919). Brighton, UK: IEEE. doi:10.1109/ICASSP.2019.8683667

20. Kurian, R. E., & Al-Assaf, Y. (2020). Impact of high school curriculum on student performance at university. (pp. 1-7). Seattle, WA, USA: IEEE Global Humanitarian Technology Conference (GHTC). doi:https://doi.org/10.1109/GHTC46280.2020.9342924

21. Latif, A., Choudhary, A., & Hammayun, A. (2015). Economic Effects of Student Dropouts: A Comparative Study. *Journal of Global Economics, 3*(2), 137. doi:10.4172/2375-4389.1000137

22. Lu, S.-X., & Wang, X.-Z. (2004). A comparison among four SVM classification methods: LSVM, NLSVM, SSVM and NSVM. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics. 7*, pp. 4277-4282. Shanghai: IEEE Cat. No.04EX826. doi:https://doi.org/10.1109/ICMLC.2004.1384589

23. Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., . . . Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048-3061. doi:10.1109/TKDE.2019.2962680.

24. McIntosh, A. M., Sharpe, M., & Lawrie, S. M. (2010). *Companion to Psychiatric Studies: Research methods, statistics and evidence-based practice.* Churchill Livingstone: Elsevier. doi:https://doi.org/10.1016/B978-0-7020-3137-3.00009-7

25. Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning

Approach: A Case of Tanzanian’s Secondary Schools. *Applied artificial intelligence, 36*(1), 451. doi:https://doi.org/10.1080/08839514.2022.2071406

26. Muschelli, J. (2020, October). ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification, 37*(3), 696-708. doi:10.1007/s00357-019-09345-1
27. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *https://www.sciencedirect.com/journal/computers-and-education-artificial-intelligence, 3*, 100066. doi:https://doi.org/10.1016/j.caeai.2022.100066
28. OECD. (2022). *Education at a Glance 2022: OECD Indicators.* Paris: OECD Publishing. doi:https://doi.org/10.1787/3197152b-en
29. Powers, D. M. (2020, October 11). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *ProQuest, 1*, 2331-8422. Retrieved from http://arxiv.org/abs/2010.16061
30. Riquelme, N., Lücken, C. V., & Baran, B. (2015). Performance metrics in multi-objective optimization. *2015 Latin American Computing Conference (CLEI)* (pp. 1-11). Arequipa, Peru: IEEE. doi:10.1109/CLEI.2015.7360024
31. Saltelli, A., Tarantola, S., F., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice- A Guide to Assessing Scientific Models.* New Jersey, US: John Wiley & sons.
32. Stehlik-Barry, K., & Babinec, A. J. (2017). *Data Analysis with IBM SPSS Statistics : Master Data Management and Analysis Techniques with IBM SPSS Statistics.* Birmingham, UK: Packt Publishing, Limited. Retrieved from https://ebookcentral.proquest.com/lib/rit/detail.action?docID=5058271
33. Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining* (Pearson New International edition. First edition. ed.). Pearson: Pearson. Retrieved from https://go.exlibris.link/Q652r2vC
34. Willcoxson, L., Cotter, J., & Joy, S. (2011). Beyond the first-year experience: the impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities. *36*(3), 331-352. doi:https://doi.org/10.1080/03075070903581533
35. Zang, L., & Rangwala, H. (2018). Early Identification of At-Risk Students Using Iterative Logistic Regression. *Artificial Intelligence in Education 19th International Conference* (pp. 613-626). London, UK: Springer International Publishing. doi:10.1007/978-3-319-93843-1_45