

Rochester Institute of Technology

RIT Digital Institutional Repository

Theses

5-21-2024

Using Prediction ML algorithm for predicting early Student Attrition in Higher Education

Noora Ali Mohsen AlAttar AlHashmi
naa6301@rit.edu

Follow this and additional works at: <https://repository.rit.edu/theses>

Recommended Citation

AlHashmi, Noora Ali Mohsen AlAttar, "Using Prediction ML algorithm for predicting early Student Attrition in Higher Education" (2024). Thesis. Rochester Institute of Technology. Accessed from

This Thesis is brought to you for free and open access by the RIT Libraries. For more information, please contact repository@rit.edu.

Using Prediction ML algorithm for predicting early Student Attrition in Higher Education

by

Noora Ali Mohsen AlAttar AlHashmi

**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 & Research

Rochester Institute of Technology Dubai

**Graduation Date
21 May 2024**

RIT

Master of Science in Professional Studies: Data Analytics

Graduate Thesis Approval

Student Name: **Noora Ali Mohsen AlAttar AlHashmi**

Thesis Title: **Using Prediction ML algorithm for predicting early Student Attrition in Higher Education**

Graduate Committee:

Name: **Dr. Sanjay Modak**

Date:

Chair of Committee

Name: **Dr. Hammou Messatfa**

Date:

Member of Committee:

Acknowledgments

I would want to describe my sincere appreciation toward Dr. Sanjay (Assistant Professor and Chair) all the instructors at RIT and special thanks to my mentor, Dr. Hammou Messatfa, for his advice and assistance throughout my thesis. I would certainly also want to appreciate my fellow classmates for their encouraging comments and words of advice throughout my studies. Despite my little experience of programming languages like as R, SPSS and Python, I value every piece of information I got through Rochester Institute of Technology Dubai, as well as the replies to my inquiries from each teacher. Last but not least, I want to thank family as well as friends for supporting me and encouraging me to reach my goals.

Abstract

This study aims to use predictive models in order to figure out students that are in likelihood of attrition as well as determine the factors that may contribute to this attrition. Student attrition is the occurrence in which students leave the institution without finishing their studies or taking the degree. It includes students either willingly or involuntarily terminating their education and failing to graduate. The findings will enable relevant parties to establish successful strategies or approaches and activities to assist in reducing the number of students who leave. With the testing and analysis approaches, it was recognized that SVM 1 was the most effective algorithm holding a 90.928% accuracy rate with precision 0.921. Students attrition of classes for a variety of motives including personal or academic challenges that prevent them from becoming engaged students. The paper seeks to investigate the numerous factors that impact the decision of the students to attrite, as well as the optimal prediction model. Kaggle will be used as a source for the data and SPSS will be used for examining and preprocessing the chosen data. Also, through SPSS Modeler will additionally be used to extract visual insights using the provided dataset.

Keywords: Higher Education, Attrition, Students, Attrition Risk, Higher Education Institutions, Return, Machine Learning Techniques, Regression, Principal Component Analysis, SVM 1, Random Trees 1, Logistic regression 1, Neural Net 1, CHAID 1 and XGBoost Tree 1.

Table of Contents

ACKNOWLEDGMENTS	II
ABSTRACT	III
LIST OF FIGURES	VI
LIST OF TABLES	VII
CHAPTER 1	1
1.1 INTRODUCTION	1
1.2 BACKGROUND INFORMATION	1
1.3 PROBLEM STATEMENT	3
1.4 RESEARCH AIM AND OBJECTIVES	4
1.5 RESEARCH QUESTIONS	4
1.6 LIMITATIONS OF THE STUDY	4
1.7 STRUCTURE OF THE THESIS	5
CHAPTER 2 LITERATURE REVIEW	6
2.1 LITERATURE REVIEW KEY TAKEAWAYS	13
CHAPTER 3 RESEARCH METHODOLOGY	14
CHAPTER 4 FINDINGS AND DATA ANALYSIS	15
4.1 DATA USED:	15
4.1.1 DATA COLLECTION	15
4.1.2 DATASET INFORMATION	16
4.1.3 VARIABLE DICTIONARY	17
4.2 EXPLORATORY DATA ANALYSIS	18
4.2.1 DATA QUALITY	19
4.2.2 DATA CLEANING	20
4.2.3 DIMENSIONS REDUCTION AND FEATURE ENGINEERING	21
4.2.4 DATA VISUALIZATION	27
4.3 MACHINE LEARNING MODEL DEVELOPMENT:	35
CHAPTER 5 DISCUSSION	51
CHAPTER 6 CONCLUSIONS & FUTURE WORK	54
6.1 CONCLUSION	54
6.2 RECOMMENDATIONS	56
6.3 FUTURE WORK	56

REFERENCES.....	58
APPENDIX.....	62
1. SCENARIO 0:.....	62
2. SCENARIO 1:.....	64
3. SCENARIO 2:.....	66

List of Figures

Figure 1. Dataset Variables Description	16
Figure 2. PCA Inputs	22
Figure 3. The Distribution of DEGREE_GROUP_DESC	27
Figure 4. Distribution of CORE_COURSE_NAME_2_F.....	28
Figure 5. (STDNT_AGE) Variable Missing Data.....	29
Figure 6. (DISTANCE_FROM_HOME) Variable Missing Data	29
Figure 7. (HIGH_SCHL_GPA) Variable Missing Data.....	30
Figure 8. The New Group Variable Named as (New_Age)	31
Figure 9. Boxplot of Variables that Have Impacted on the Return of The Students.....	33
Figure 10. Showing the Proportion of the Student’s Return Before the SMOTE Technique	38
Figure 11. Showing Balance Directives Factors of Student’s Return During SMOTE Technique.....	39
Figure 12. Showing the Proportion of the Student’s Return After the SMOTE Technique.....	39
Figure 13. Confusion Matrix Explaining the Predicted and the Actual	40
Figure 14. Predictor Importance for SVM 1 in Scenario Three	45
Figure 15. The Distributions of the Variable SECOND_TERM_ATTEMPT_HRS	47
Figure 16. The Distributions of the Variable HIGH_SCHL_GPA.....	47
Figure 17. The Distributions of the Variable FIRST_TERM_EARNED_HRS.....	47
Figure 18. The Distributions of the Variable Factor 1.....	47
Figure 19. The Distributions of the Variable Factor 6.....	48
Figure 20. P-Value < 0.001 of The Chi-Square of the Relation	49
Figure 21. ROC for SVM 1 in Scenario Three	50
Figure 22. List of Models Used for Scenario Zero.....	63
Figure 23. Predictor Importance for Random Tree 1 in Scenario Zero.....	63
Figure 24. ROC for Random Tree 1 in Scenario Zero	64
Figure 25. List of Models Used for Scenario One.....	65
Figure 26. Predictor Importance for LVSM 1 in Scenario One.....	65
Figure 27. ROC for LVSM 1 in Scenario One	66
Figure 28. List of Models Used for Scenario Two	67
Figure 29. Predictor Importance for Random Tree 1 in Scenario Two.....	67
Figure 30. ROC for Random Tree in Scenario Two.....	68

List of Tables

Table 1. Dataset Variables Description	18
Table 2. Showing Before the Null Values of Variables are 100% Completed.....	19
Table 3. Showing The Amputation Method Used for the Uncompleted Variables.....	20
Table 4. Showing After the Null Values of Variables are 100% Completed	21
Table 5. Using the PCA Methods for 12 Variables	26
Table 6. Student New Age Group Classification.....	31
Table 7. Shows The Importance of Each Variable That Have Impact on the Return of The Students	32
Table 8. Showing The 22 Important Variables That are Used in Order to Build Up he Model.....	36
Table 9. List of Variables Used for Scenario Three	36
Table 10. List of Models used for Scenario Three	43
Table 11. Hypothesis Test Summary Using Mann-Whitney	46
Table 12. P-value of the null hypothesis using the Chi-Square.....	49
Table 13. Variables Used for Scenario Zero.....	62
Table 14. Variables used for Scenario One	64
Table 15. List of Variables used for Scenario Two	66
Table 16. Demonstrating All the Scenarios Used While Building the Models	68

Chapter 1

1.1 Introduction

Student attrition, often known as withdrawal or dropout, that is a serious issue in the world of education. It refers to the phenomena in which students attrite or fail to finish their studies. This mostly happen at the stage of higher education, and is impacted by a variety of aspects including professional, personal, and economical conditions. Recognizing and solving student attrition is critical for educational organizations and legislators who want to guarantee that students may fulfill their educational objectives and attain their maximum potential.

1.2 Background information

In general, the rise in the rates of student attrition among the higher education have damaged both the institutions and individuals negatively. Because of online learning and the pandemic in the past few years, it is now critical for institutions to determine the indicators that impact students' withdrawal from institutions of higher education. The consequences of student attrition have affected the educational institutions' revenue and reputation, it has turned into an unacceptable risk that has forced them to investigate the causes and make solutions to resolve this problem (Garrison, 2017).

Previous research on student attrition rates, including the various models utilized for earlier prediction has been examined in the present paper. As a result of the findings, it appears that several variables influence students' decisions about continuing their education at the college or not, allowing institutions to establish methods that tackle this problem as it arises.

Many people believe that withdrawn students are poor achievers, fail frequently, or have significant absence rates. Nevertheless, this could not be the case because there may be non-academic issues that impact the decision to leave the university. On the basis of the used data, a predictive model is going to be developed for important stakeholders in order to utilize in generating corrective action plans for supporting in reducing attrition rates. The data contains of around 8,000 data of under-graduated students from a large popular public institution was used to train machine learning algorithms in order to predict the issue of student attrition. A number of variables such as, socioeconomic, academic and demographic as that of predictors, which might involve gender, GPA, age, SAT scores, major and family income.

It is mainly important to know the number of graduates from each university, as this quantitative value represents the success percentage for education providers. In 2018, Sub-Saharan Africa led the world in student attrition rates at 37.5%, preceded by South Asia starting at 15.5% as well as the Middle East at around 11% (Stastica 2022). According to Ahmed, Al-Mansoori, Khan, and Hassan (2020) discovered that 12% is the attrition rate within a selected institution, with greater rates reported in various fields along with students representing different socioeconomic backgrounds. However, it is crucial to recognize students who are willing to attire and the causes behind this behavior impacting the business of education (Seltzer, 2021). As pointed out by Bean & Metzner (1985), the business impact concerning student attrition consists of missed tuition payments, the reputation of the institute and branding damage, and the possibility of a decline in total enrollment.

Winters, Xu, and Guo (2019), claimed that universities are affected severely when it comes to financial expenses because of numerous causes for students leaving out since it costs around \$6,500 and \$12,300. Hence, eliminating the rates of attrition within 1% might enhance university income by at least \$1.1 million per year. On the other hand, the study of Zhang & Johnes (2017), resulted in

finding out that attrition has an important effect on student success such as the rates of degree completion, grade average, as well as employment outcomes. The existence of attrition may influence these figures and as a result, it will eventually affect the economy and society. Therefore, it is a crucial matter to discover what are the main factors and contributors that support students to leave their higher education at an early stage.

1.3 Problem Statement

Recently, students are dropping out from higher education at their entry level that leaves higher educational institutes in an unknown situation.

The term "student attrition" refers to a decline in the overall number of studying students in a college or university (Higher Education). The student's records or numbers who withdraw from their courses prior to graduation are recorded. As pupil leave their higher education, many problems impact the education institutions and end with some institutional consequences without being able to point out the most significant and critical issue behind student attrition causing these negative consequences to the universities. Some organizations use attrition rate to gauge the success of educational institutions. Besides, several studies show that the educational sector is trying their best efforts in order to understand the challenges behind students dropping out and reducing their attrition level.

1.4 Research Aim and Objectives

This paper has two primary objectives: (a) Determine the features that impact students' decision to leave a university. (b) Examine models that can properly forecast which students have a risk of attrition. Furthermore, I will read other journal papers to gain insight into the various techniques employed by the writers to address this topic and the limits of their research.

1.5 Research Questions

1. Assess the machine learning models effectiveness to predict student's attrition.
2. Enhance quality of the dataset through data preprocessing techniques and cleaning.
3. Implement feature selection methodology to highlight the main key predictors of student attrition.

1.6 Limitations of the study

While completing this work, there had been just a few constraints:

- **Data ownership:** I got the data from Kaggle.
- **Governance of the model and accountability:** For the built model, we did not have a chance of deploying it, which considered to be with unpredictable behavior and there is less chance of knowing how the model will behave after a certain period of time.

- **Maintenance and monitoring:** The model needs continuous monitoring that will support in detecting and addressing the performance degradation.
- **Feedback loop:** Collecting the feedback from the teachers or the universities has complex process specially in the reinforcement of the learning scenarios.

1.7 Structure of the thesis

Chapter2: Highlights the theories of the field, as well as the prediction techniques, difficulties, and gaps.

Chapter 3: The research approach is presented in Chapter 3. It also includes an overview of the steps involved in preparing, separating, as well as data calling before going on to discuss the algorithms which will be applied and contrasted.

Chapter 4: In-depth explanations of the results and analysis of the data procedure are provided in the chapter along with methods for data preparation, prediction model construction, including exploratory data analysis. This also provides an example of the dashboard visualization.

Chapter 5: The research discoveries are discussed as well as the research questions are addressed in light of the findings.

Chapter 6: provides an overview of the research methodology and findings, highlighting its limits and offering recommendations for further study.

Chapter 2 Literature Review

Attrition from higher education has been linked to a wide range of issues. Several journal articles address the problem of student attrition and the factors that affect the student's decisions. Also, I will examine the various prediction methods that the authors utilized in their articles in order to identify the shortcomings of their research and attempt to come up with a solution that will improve on previous studies.

Berens et al (2019) created an Early Detection System (EDS) utilizing data collected from private as well as public colleges to successfully forecast student attrition. Rather than depending just on one approach, the models have been constructed using a variety of techniques, including AdaBoost Algorithm, which included decision trees, regression analysis as well as networks.

Accuracy prediction was performed in two stages, at first-semester end and at the end of fourth semester. The outcomes demonstrated that data accuracy rises with time. This means that in semester four the accuracy rises whenever the model estimates the accuracy. The disadvantage is the fact that the current demographic data can only be useful for early diagnosis in year one since, by that time, the disease has progressed.

The introduction in algorithmic fairness is a requirement, which maintains that decisions made by automated systems ought to be fair in terms of attributes that are protected such as (gender, race, etc). Protective feature bias as well as class imbalance may both be identified within training datasets. We hypothesize that in order to improve model accuracy as well fairness, it is necessary to decrease bias in the two protected features and classes. Using SMOTE reduces class imbalance as

well as improves group fairness by enhancing feature blurring (Dablain, D., Krawczyk, B., & Chawla, N. V. 2022).

Thammasiri, & Kasap (2014) clarified that for balancing the data, they have oversampled, and then they employed three different classifiers for evaluation as well as prediction. Models contained logistic regression, decision trees, as well as support vector-machines. The findings indicate that when it comes to identifying students who are at threat of attrition logistic regression and SVM algorithms outperform the decision tree when it comes to classification accuracy.

In regards of feature selection, it has numerous potential advantages, including lowering measurement, storage needs, speeding up training and overcoming the dimensionality to enhance prediction performance. Also, it supports in enabling data visualization as well as comprehension (Guyon and Elisseeff, 2003; Saeys et al., 2007).

Abu Oda & ElHalees (2015) aimed to determine students who are more unlikely to make it back to the program of computer science through one semester into the next in a journal publication. The researchers used information collected by Al-Aqsa University among students studying bachelor, including 1290 entries representing students along with transcripts. During the study, researchers employed several categorization algorithms to forecast and assess student attrition. Decision-tree (DT) as well as Nave Bayes (NB) approaches are among them. The findings indicated how mastering courses that include algorithm analysis and digital design has a significant impact on students' program retention prediction and decreasing attrition rates.

Hirschy & McClendon (2004) stated that identifying the factors that cause college student attrite is critical for designing successful retention measures. A higher student engagement rate, participation in campus events, along with assistance from professors and staff may have been discovered in the

study to lead to greater retention rates plus lower student departure. Academic services, including tutoring, counseling, and mentorship programs, may have been emphasized as important in boosting student success while lowering attrition throughout the research. Also, the study may have looked at the impact of institutional regulations on student attrition rates, including admissions standards, scholarships availability along with support systems.

According to Daffertshofer & Lamoth (1997) clarified that Principal Component Analysis (PCA) is reflected as a common reduction of dimensionality method used in analysis of data. Its appeal stems from three key characteristics. Initially it is the best (in regard to mean squared-error) that is linear strategy for reducing and rebuilding a collection of highly dimensional variables through a collection of less or low dimensional variables.

Also, a study underlines the relevance of financial help in reducing attrition chances. It implies that tailored financial assistance measures can have a significant impact upon minority students' academic achievement and retention (Chen & DesJardins, 2010).

The authors Saltelli, Tarantola, Campolongo, and Ratto (2004) address the significance of sensitivity analysis of the model outputs are affected by adjustments to input variables and offer useful techniques for applying sensitivity analysis. Additionally, they illustrated that sensitivity analysis is crucial to comprehend in order to evaluate how solid their findings and conclusions are. Researchers may assess how adjustments to important assumptions particular parameters affect their study's outcomes by using sensitivity analysis. This eventually strengthens the validity as well as dependability of their findings by assisting in the identification of potential causes of bias as well as uncertainty. All things considered, sensitivity analysis comprehension is essential to guaranteeing reliable and accurate research findings. Many sensitivity analysis methods are covered also such as

variance-based methodologies, global sensitivity analysis, and local sensitivity evaluation. The authors also provide insights on the estimation of uncertainty as well as the confirmation of sensitivity analysis outcomes.

‘Spielman & Julka (2004) mentioned in his research that the researchers often examine several characteristics that may impact students' academic achievement along with the choice to continue or discontinue their studies as indicators of academic accomplishment and retention within college freshman. Variables including GPA of high school, test scores that are standardized, socioeconomic background, participation of the student, academic assistance services use, along with social integration may be taken into account.

Grau-Valldosera & Minguillón (2014) point out that distinct factors influence students' decisions to withdraw from typical on-campus courses vs online courses. Thus, they defined student attrition being enrolling at a university at some point in their lives but leaving before graduating.

A study conducted by AlJohani (2016), his findings suggests that students who are unaware of the various sorts of schools' offers in terms of educational opportunities and job chances are in danger of attrition compared to individuals who are aware. His research focused on attrition in the fourth year and second-year colleges and universities.

Shaw et al. (2016) found that students leave from online universities at a rate that is 5% greater compared to the traditional ones. As a result, they have developed a tracking mechanism to monitor the variables affecting the attrition within studying online courses as well as to determine the students impacted by such variables. This aided them in creating retention strategies to prevent at-risk students from attriting.

A study showed that when a dataset is imbalanced, it means that one class is expressively more prevalent than others, it can indicate to problems such as biased predictions, poor generalization, and lower accuracy on minority classes (Mduma, N. 2023).

Lee and Chung (2019) asserted that student attritions have an impact on society in addition to having a direct impact on students. Attrition will have an effect on the student's well-being in later years since they will lose the opportunity of gaining new information and abilities that they require in their potential career path. In terms of society, the more students that attrite, the less qualified workers there will be for its companies, which will lead to higher unemployment levels. Lee and Chung attempted to predict early student attrition employing a variety of approaches, including synthetic with trained classifiers for random-forest (RF), random forest via SMOTE(SMOTE + RF), boosted-decision tree(BDT) as well as boosted decision-tree along with SMOTE(SMOTE + BDT). Because barely any information was available for further research, the outcomes were not particularly helpful. As a result, the authors noted that additional research into student attrition caused by class imbalance is required in the future.

According to Ananat, Gassman-Pines, and Gibson-Davis (2013), there are many economic impacts of student attrition such as losing the education cost that the students received up to that time is wasted, the attrited students lose the opportunity of getting highly paid jobs and they will miss the chance of learning significant skills which can be used in innovation.

Persson and Rossin-Slater (2018), stated that attriting school and unemployment at an early age may result in serious long-term consequences for illnesses. Furthermore, it might have a negative effect on social mobility as well as higher crime rates. Recognizing the long-term effects of student

attrition and striving to avoid it may represent a significant step toward achieving social well-being along with equity.

According to Maher & Macallister (2013) mentioned in their published journal paper, there are several extant publications that discussed the various implications for the topic attrition of student in institutes of higher education. These consequences are not only on the institutions, though they impact the students as well as staff on personal level. The major purpose of their article is to identify the most important features affecting student retention that can be utilized in institutes of higher education. The authors employed a combination of ways to collect data, including data regarding the issue of student attrition along with retention rates, as well as interviews involving relevant staff. Their research concludes that once students attrite both institutions and individuals suffer financial losses.

Johnson (2012) argues in a journal paper that there have been financial effects upon the institution of higher education owing to student attrition. As a result his data indicate that students around 35% abandon their studies before completing the whole academic plan. The author emphasizes that the consequences for colleges are identical regardless the students elect to move to other less costly schools or abandon their studies completely. As noted by Johnson, colleges spend around \$43K for every student plus roughly \$18K for those who withdraw, indicating that their resources are not used to their maximum potential, putting the institution or colleges at a loss of money with no way out for profits.

A study by Pascarella and Terenzini (1980) declared that, on forecasting voluntary first-year persistence as well as withdrawal behavior within a domestic institution. This was with a focus on Tinto's model validation while having the influence of different factors such as student experience,

institutional factors, and pre-college characteristics. The researchers examined the correlations between numerous factors within Tinto's model as well as their influence on freshman-year attrition or using path analysis, which is a statistical approach.

Aulck et al. (2016) reported that within United States the higher education, where even more than \$ 9 billion was invested in educating these students, about 30% of those first year students did not return for studying in their second-year.

The biggest dataset concerning higher education was utilized by the researchers to analyze student academic performance and demographics from of the largest colleges. Even though the dataset only covered one term, they were able to utilize it to develop a model for predicting student attrition rates. The major indicators of student attrition, according to the findings, were English, chemistry, math, and psychology.

They have employed a variety of techniques, including such as, experimental, quantitative as well as, and correlational design methodologies, to identify the crucial features related to students' failure to complete their study results. According to the research, students who take courses that emphasize verbal learning seem to be more likely to attrite compared to the students who take courses that focus on skill development. However, students who attrite for personal purposes in order to resume them the next term are additionally more likely to start attriting. Therefore, for advice, the authors suggested using the tracking system that is created to identify students who might leave so that additional support activities might be suggested for them to assist them deal with the challenges.

We need to test various models and compare them using various metrics to support in determining the categorization model that best forecasts student attrition. Following the development of the optimal model, we must periodically assess it because data can change. In

addition, there seem to be good techniques for classification, including Weka, that are utilized by teachers in addition to machine learning instruments like SPSS, R Studio, and Python. The data being accessible, developing methods for dealing with the enormous quantity of data , and determining an appropriate model that had high reliability were all factors that limited the study.

2.1 Literature Review Key Takeaways

- I have learnt from the literature review, the importance of identifying the factors of student attrition to design effective retention measures.
- I have reached into a conclusion that online universities are 5% higher in attrition than traditional ones. Also, I knew from this study that I can develop a tracking mechanism to monitor the variables affecting attrition.
- I have realized that there are many techniques could be used such as, decision trees (DT) and Naive Bayes (NB) algorithms to identify student attrition.
- I have resulted from the literature review that student attritions have an impact on society as well as student's wellbeing.
- I have absorbed from the literature review that the economical impacts can be resulted from student attrition such as, loosing the education cost and the attrited students loose the opportunity of getting highly paid jobs.
- I have reached into a conclusion that attriting school may result in illnesses and negatively impact social mobility and higher crime rate.

Chapter 3 Research Methodology

To attain the greatest outcomes, the Cross-Industry Standard Process for Data-Mining “CRISP-DM” technique, as described further down is the ideal way to proceed for the project. This technique supports in implementing the findings in a very organized and structured project. It is an exceptionally used framework for carrying out data science tasks. It describes a data-science framework (the process in data-oriented projects) in a logical way.

The technique can be simplified in six phases business. The first phase is business understanding which concentrates on identifying the project goals and needs through a business standpoint, afterwards translating this information through a data mining issue definition plus an initial plan for the project tailored to meet the objectives.

The second phase is the data understanding that involves data collection along with actions to become acquainted through the data, find data quality issues, uncover early discoveries about the data, while curious subsets to create hypotheses about undiscovered or hidden data. The following phase is data preparation involves in encompassing all operations that result in the final set of data of which will be supplied through modeling techniques being constructed beyond the original data. In the next phase which is modeling, several modeling approaches are chosen and employed in this stage, and the parameters they require are adjusted to ideal levels. Evaluation stage, in this stage several models which seem to be superior quality according to the analysis of the data.

The last phase is development, in most cases, the model development is not the final step of the work in progress. Typically, the acquired knowledge must be structured and delivered in an approach that the end user can use.

Chapter 4 Findings and Data Analysis

4.1 Data Used:

The project will require going through numerous processes to determine the optimum model in order to be used with the obtained dataset. Initially the dataset has to pass via a preprocessing stage during which it will be changed using raw data into a format which can enable us to comprehend it. Finding information and working with various tools as well as programming languages for obtaining insights constitutes a crucial stage within data mining. During this step, data cleaning strategies that involve refilling null values shall be used. Additionally, rather than dealing with all qualities, choose the ones that will help us succeed in this project. Whenever the dataset is organized, cleaned and suitable for analysis, the following step will be gaining insights using visualizations.

4.1.1 Data Collection

Several publicly accessible databases were investigated in order to locate the appropriate dataset that could be used with, which includes Dubai Pulse, Data.Gov, USA Census-Bureau, that is regarded as the trove treasure for US data, as well as finally, UNICEF Dataset. Despite this, the proposed dataset for this study is taken from Kaggle website under the title (Students' Early Attrition Analysis).

<https://www.kaggle.com/vijaysimhan/student-admissions-data-for-a-university>

4.1.2 Dataset Information

The dataset contains old historical data about students who left their higher education and University details. The dataset includes some demographic information such as type of students, grades, course details, and financial aspects.

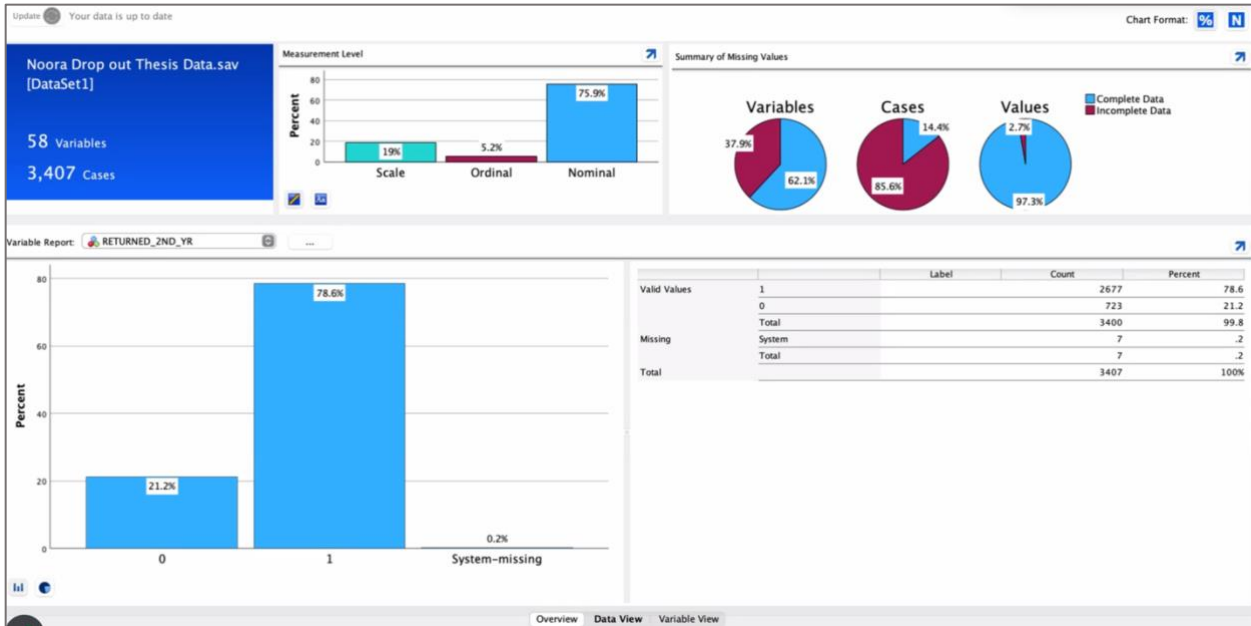


Figure 1. Dataset Variables Description

4.1.3 Variable Dictionary

The open data contains 58 attributes as shown in the below table. Categorical and 3400 continuous records, around 33 outliers shall be considered in the dataset. Also, some null values will be tackled and cleaned during the pre-processing stage.

Variable Name	Description
STUDENT IDENTIFIER	Student Identifier
STDNT_AGE	Age of the Student Enrolled
STDNT_GENDER	Gender of the student
STDNT_BACKGROUND	Background of Student
IN_STATE_FLAG	Indicator of whether Student is in the same state as the university
INTERNATIONAL_STS	Indicator of whether Student is an International Student
STDNT_MAJOR	Student's Major course in university
STDNT_MINOR	Student's Minor course in university
STDNT_TEST_ENTRANCE1	Student's Entrance 1 score
STDNT_TEST_ENTRANCE2	Student's Entrance 2 score
STDNT_TEST_ENTRANCE_COMB	Student's score calculated both on Entrance1 & Entrance2 score
FIRST_TERM	First semester year
CORE_COURSE_NAME_1_F	Core course 1 opted in the First semester
CORE_COURSE_GRADE_1_F	Grade in Core course 1 opted in the First semester
CORE_COURSE_NAME_2_F	Core course 2 opted in the First semester
CORE_COURSE_GRADE_2_F	Grade in Core course 2 opted in the First semester
CORE_COURSE_NAME_3_F	Core course 3 opted in the First semester
CORE_COURSE_GRADE_3_F	Grade in Core course 3 opted in the First semester
CORE_COURSE_NAME_4_F	Core course 4 opted in the First semester
CORE_COURSE_GRADE_4_F	Grade in Core course 4 opted in the First semester
CORE_COURSE_NAME_5_F	Core course 5 opted in the First semester
CORE_COURSE_GRADE_5_F	Grade in Core course 5 opted in the First semester
CORE_COURSE_NAME_6_F	Core course 6 opted in the First semester
CORE_COURSE_GRADE_6_F	Grade in Core course 6 opted in the First semester
SECOND_TERM	Second semester year
CORE_COURSE_NAME_1_S	Core course 1 opted in the Second semester
CORE_COURSE_GRADE_1_S	Grade in Core course 1 opted in the Second semester
CORE_COURSE_NAME_2_S	Core course 2 opted in the Second semester
CORE_COURSE_GRADE_2_S	Grade in Core course 2 opted in the Second semester
CORE_COURSE_NAME_3_S	Core course 3 opted in the Second semester

CORE_COURSE_GRADE_3_S	Grade in Core course 3 opted in the Second semester
CORE_COURSE_NAME_4_S	Core course 4 opted in the Second semester
CORE_COURSE_GRADE_4_S	Grade in Core course 4 opted in the Second semester
CORE_COURSE_NAME_5_S	Core course 5 opted in the Second semester
CORE_COURSE_GRADE_5_S	Grade in Core course 5 opted in the Second semester
CORE_COURSE_NAME_6_S	Core course 6 opted in the Second semester
CORE_COURSE_GRADE_6_S	Grade in Core course 6 opted in the Second semester

Table 1. Dataset Variables Description

4.2 Exploratory Data Analysis

Loading required SPSS modeler in order to handle the dataset. It's the first thing to do is to examine the dataset's most fundamental information, including the number of variables there are, whether there are null values, the names of the columns, and dataset overview.

Throughout the dataset, presently are several sorts of properties that include integers, floats, as well as objects. That is additionally obvious which there are many null values for each attribute; within this scenario, the data must be extracted to ensure quality.

A few attributes, including fundamental course titles along with grades, contain the highest null values percentages, with some reaching 97%. A number of methods for filling with null values which are going to be discussed in the following section.

Data distribution and correlations between variables is an important aspect to understand data in depth. A wide exploratory data analysis was employed to predict early student attrition within higher education firms.

4.2.1 Data Quality

As observed in previous part, the dataset contains a large number concerning null values. In order to be capable to function using the data that we have, data must be cleansed and filled within null values. For the attribute with a distribution that is normal, null values have been filled with mean. For as some other attributes, CRT have been used in order to fill the null values. In which, only 72.24% of the records are completed and 67.65% are completed fields therefor the rest requires to be completed.

However, below are the variables that requires to be completed to 100% which are:

CORE_COURSE_NAME_3_F, CORE_COURSE_GRADE_3_F, CORE_COURSE_GRADE_2_S,
 CORE_COURSE_GRADE_1_S, SECOND_TERM_EARNED_HRS,
 SECOND_TERM_ATTEMPT_HRS, CORE_COURSE_NAME_1_S,
 CORE_COURSE_NAME_2_F, CORE_COURSE_GRADE_2_F, HIGH_SCHL_GPA and
 DISTANCE_FROM_HOME.

Audit Quality Annotations										
Complete fields (%): 67.65% Complete records (%): 72.24%										
Field	Measurement	Impute Missi...	Method	% Complete /	Valid Records	Null Value	Empty String	White Space	Blank Value	
▲ CORE_COURSE_NAME_3_F	▲ Nominal	... Never	Fixed	83.382	2835	0	565	565	0	0
▲ CORE_COURSE_GRADE_3_F	▲ Nominal	... Never	Fixed	83.382	2835	0	565	565	0	0
▲ CORE_COURSE_GRADE_3_S	▲ Nominal	... Never	Fixed	87.088	2961	0	439	439	0	0
▲ CORE_COURSE_GRADE_1_S	▲ Nominal	... Never	Fixed	93.176	3168	0	232	232	0	0
⊗ SECOND_TERM_EARNED_HRS	⊗ Continuous	0 ... Never	Fixed	93.853	3191	209	0	0	0	0
⊗ SECOND_TERM_ATTEMPT_H...	⊗ Continuous	4 ... Never	Fixed	93.941	3194	206	0	0	0	0
▲ CORE_COURSE_NAME_1_S	▲ Nominal	... Never	Fixed	95.382	3243	0	157	157	0	0
▲ CORE_COURSE_NAME_2_F	▲ Nominal	... Never	Fixed	97.088	3301	0	99	99	0	0
▲ CORE_COURSE_GRADE_2_F	▲ Nominal	... Never	Fixed	97.088	3301	0	99	99	0	0
⊗ HIGH_SCHL_GPA	⊗ Continuous	0 1 ... Never	Fixed	98.441	3347	53	0	0	0	0
⊗ DISTANCE_FROM_HOME	⊗ Ordinal	... Never	Fixed	99.265	3375	25	0	0	0	0
⊗ STUDENT_IDENTIFIER	⊗ Continuous	0 0 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ STDNT_AGE	⊗ Ordinal	... Never	Fixed	100	3400	0	0	0	0	0
▲ STDNT_GENDER	▲ Flag	... Never	Fixed	100	3400	0	0	0	0	0
▲ STDNT_BACKGROUND	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ IN_STATE_FLAG	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ STDNT_MAJOR	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ STDNT_MINOR	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
⊗ FIRST_TERM	⊗ Ordinal	... Never	Fixed	100	3400	0	0	0	0	0
▲ CORE_COURSE_NAME_1_F	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ CORE_COURSE_GRADE_1_F	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
⊗ SECOND_TERM	⊗ Ordinal	... Never	Fixed	100	3400	0	0	0	0	0
▲ HOUSING_STS	▲ Flag	... Never	Fixed	100	3400	0	0	0	0	0
⊗ RETURNED_2ND_YR	⊗ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ FATHER_HI_EDU_DESC	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ MOTHER_HI_EDU_DESC	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ DEGREE_GROUP_CD	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
▲ DEGREE_GROUP_DESC	▲ Nominal	... Never	Fixed	100	3400	0	0	0	0	0
⊗ FIRST_TERM_ATTEMPT_HRS	⊗ Continuous	0 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ FIRST_TERM_EARNED_HRS	⊗ Continuous	0 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ GROSS_FIN_NEED	⊗ Continuous	2 0 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ COST_OF_ATTEND	⊗ Continuous	0 0 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ EST_FAM_CONTRIBUTION	⊗ Continuous	29 ... Never	Fixed	100	3400	0	0	0	0	0
⊗ UNMET_NEED	⊗ Continuous	2 ... Never	Fixed	100	3400	0	0	0	0	0

Table 2. Showing Before the Null Values of Variables are 100% Completed

4.2.2 Data Cleaning

As observed in previous part, the dataset contains a large number concerning null values. In order to be capable to function using the data that we have, data must be cleansed and filled within null values. As demonstrated by the above table, the following 11 variables need to be imputed to 100% which are: CORE_COURSE_NAME_3_F, CORE_COURSE_GRADE_3_F, CORE_COURSE_GRADE_2_S, CORE_COURSE_GRADE_1_S, SECOND_TERM_EARNED_HRS, SECOND_TERM_ATTEMPT_HRS, CORE_COURSE_NAME_1_S, CORE_COURSE_NAME_2_F, CORE_COURSE_GRADE_2_F, HIGH_SCHL_GPA and DISTANCE_FROM_HOME.

After analyzing these missing values, we choose two techniques, one is the replacement of the missing values by the mean and the other one is using CRT algorithm.

Variable Name	Amputation Method
CORE_COURSE_NAME_3_F	Replaced the missing value by 'Unknown'
CORE_COURSE_GRADE_3_F	Replaced the missing value by 'Unknown'
CORE_COURSE_GRADE_2_S	Replaced the missing value by 'Unknown'
CORE_COURSE_GRADE_1_S	Replaced the missing value by 'Unknown'
SECOND_TERM_EARNED_HRS	Replaced the missing value using CRT Algorithm
SECOND_TERM_ATTEMPT_HRS	Replaced the missing value using CRT Algorithm
CORE_COURSE_NAME_1_S	Replaced the missing value by 'ENGL 1102'
CORE_COURSE_NAME_2_F	Replaced the missing value by 'ENGL 1101'
CORE_COURSE_GRADE_2_F	Replaced the missing value by 'B'
HIGH_SCHL_GPA	Replaced the missing value by '3.203'
DISTANCE_FROM_HOME	Replaced the missing value by '69.0'

Table 3. Showing The Amputation Method Used for the Uncompleted Variables

The above table is demonstrating the amputation method that was used in order to complete the missing variables in which two variables were replaced using the CRT algorithm and the rest of the variables were replaced by certain values.

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1	STUDENT IDENTIFIER	Continuous	0	0	None	Never	Fixed	100.000	3384	0	0	0	0
2	STDNT_AGE	Ordinal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
3	STDNT_GENDER	Flag	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
4	STDNT_BACKGROUND	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
5	IN_STATE_FLAG	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
6	STDNT_MAJOR	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
7	STDNT_MINOR	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
8	FIRST_TERM	Ordinal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
9	CORE_COURSE_NAME_1_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
10	CORE_COURSE_GRADE_1_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
11	CORE_COURSE_NAME_2_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
12	CORE_COURSE_GRADE_2_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
13	CORE_COURSE_NAME_3_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
14	CORE_COURSE_GRADE_3_F	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
15	SECOND_TERM	Ordinal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
16	CORE_COURSE_NAME_1_S	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
17	CORE_COURSE_GRADE_1_S	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
18	CORE_COURSE_GRADE_2_S	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
19	HOUSING_STS	Flag	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
20	RETURNED_2ND_YR	Flag	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
21	DISTANCE_FROM_HOME	Ordinal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
22	HIGH_SCHL_GPA	Continuous	0	1	None	Never	Fixed	100.000	3384	0	0	0	0
23	FATHER_HI_EDU_DESC	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
24	MOTHER_HI_EDU_DESC	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
25	DEGREE_GROUP_CD	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
26	DEGREE_GROUP_DESC	Nominal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0
27	FIRST_TERM_ATTEMPT_HRS	Continuous	17	0	None	Never	Fixed	100.000	3384	0	0	0	0
28	FIRST_TERM_EARNED_HRS	Continuous	54	0	None	Never	Fixed	100.000	3384	0	0	0	0
29	SECOND_TERM_ATTEMPT_HRS	Continuous	23	4	None	Never	Fixed	100.000	3384	0	0	0	0
30	SECOND_TERM_EARNED_HRS	Continuous	33	0	None	Never	Fixed	100.000	3384	0	0	0	0
31	GROSS_FIN_NEED	Continuous	2	0	None	Never	Fixed	100.000	3384	0	0	0	0
32	COST_OF_ATTEND	Continuous	0	0	None	Never	Fixed	100.000	3384	0	0	0	0
33	EST_FAM_CONTRIBUTION	Continuous	45	29	None	Never	Fixed	100.000	3384	0	0	0	0
34	UNMET_NEED	Continuous	29	2	None	Never	Fixed	100.000	3384	0	0	0	0
35	New_Age	Ordinal	--	--	--	Never	Fixed	100.000	3384	0	0	0	0

Table 4. Showing After the Null Values of Variables are 100% Completed

The above table is showing all the variables after replacing the values and using CRT algorithm to complete the missing values to 100% as shown in the table.

4.2.3 Dimensions Reduction and Feature Engineering

Principal component analysis (PCA) is considered to be a common reduction of dimensionality approach employed in analysis of data. Its appeal stems from three key characteristics. Initially it is the best (in regard to mean squared-error) that is linear strategy for reducing and rebuilding a collection of highly dimensional variables through a collection of less or low dimensional variables.

Furthermore, the parameters of the model could be calculated using the data, such as through diagonalizing covariance's of the samples. Following that, having model parameters, reduction and the process of decompression are simple processes to perform, requiring just matrix multiplications (Daffertshofer & Lamoth, 1997). The below variables have been analyzed using PCA approach.

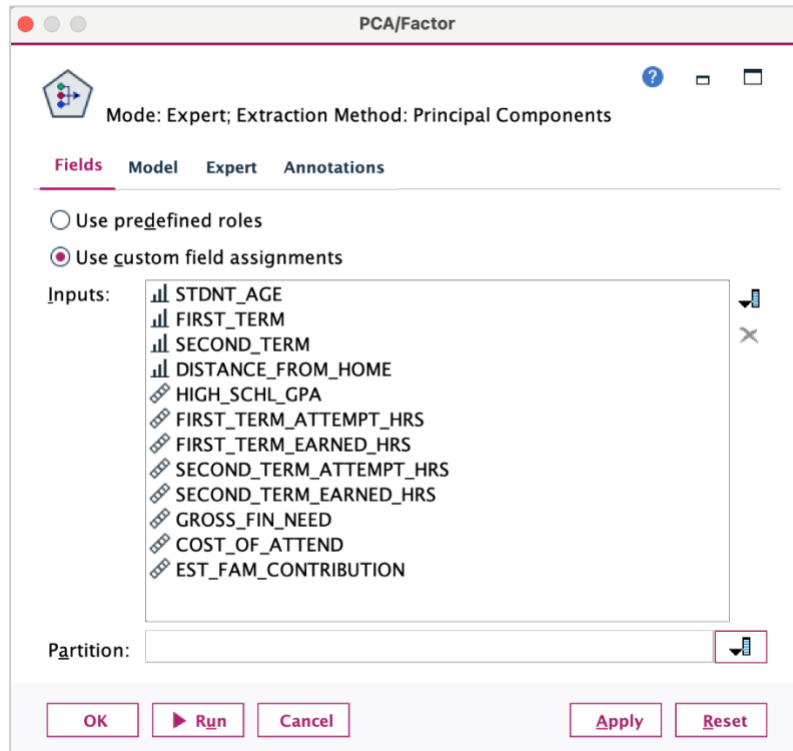


Figure 2. PCA Inputs

Below are the models for each dimension:

Equation For Factor-1

$$\begin{aligned} & -0.09378 * \text{STDNT_AGE} + \\ & -0.0006256 * \text{FIRST_TERM} + \\ & -0.0006256 * \text{SECOND_TERM} + \\ & 0.0002839 * \text{DISTANCE_FROM_HOME} + \\ & 0.4267 * \text{HIGH_SCHL_GPA} + \\ & 0.1583 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ & 0.1008 * \text{FIRST_TERM_EARNED_HRS} + \\ & 0.1359 * \text{SECOND_TERM_ATTEMPT_HRS} + \\ & 0.09323 * \text{SECOND_TERM_EARNED_HRS} + \\ & 0.0000000211 * \text{GROSS_FIN_NEED} + \\ & 0.0000001091 * \text{COST_OF_ATTEND} + \\ & 0.0000001237 * \text{EST_FAM_CONTRIBUTION} + \\ & + 245.0 \end{aligned}$$

Equation For Factor-2

$$\begin{aligned} & -0.01805 * \text{STDNT_AGE} + \\ & 0.002445 * \text{FIRST_TERM} + \\ & 0.002445 * \text{SECOND_TERM} + \\ & 0.0001459 * \text{DISTANCE_FROM_HOME} + \\ & 0.122 * \text{HIGH_SCHL_GPA} + \\ & 0.05171 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ & 0.01051 * \text{FIRST_TERM_EARNED_HRS} + \\ & 0.05957 * \text{SECOND_TERM_ATTEMPT_HRS} + \\ & 0.02814 * \text{SECOND_TERM_EARNED_HRS} + \\ & -0.0000004096 * \text{GROSS_FIN_NEED} + \\ & -0.00000039 * \text{COST_OF_ATTEND} + \\ & -0.0000001111 * \text{EST_FAM_CONTRIBUTION} + \\ & + -983.9 \end{aligned}$$

Equation For Factor-3

$$\begin{aligned} & -0.03224 * \text{STDNT_AGE} + \\ & 0.001498 * \text{FIRST_TERM} + \\ & 0.001498 * \text{SECOND_TERM} + \\ & 0.0002029 * \text{DISTANCE_FROM_HOME} + \\ & -0.07362 * \text{HIGH_SCHL_GPA} + \\ & 0.0197 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ & -0.002216 * \text{FIRST_TERM_EARNED_HRS} + \\ & 0.00663 * \text{SECOND_TERM_ATTEMPT_HRS} + \\ & -0.005608 * \text{SECOND_TERM_EARNED_HRS} + \\ & 0.0000008873 * \text{GROSS_FIN_NEED} + \end{aligned}$$

$$0.0000007906 * \text{COST_OF_ATTEND} + \\ 0.0000002226 * \text{EST_FAM_CONTRIBUTION} + \\ + -602.0$$

Equation For Factor-4

$$0.4282 * \text{STDNT_AGE} + \\ 0.00007288 * \text{FIRST_TERM} + \\ 0.00007288 * \text{SECOND_TERM} + \\ -0.001463 * \text{DISTANCE_FROM_HOME} + \\ 0.5404 * \text{HIGH_SCHL_GPA} + \\ 0.06489 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ 0.07414 * \text{FIRST_TERM_EARNED_HRS} + \\ -0.06524 * \text{SECOND_TERM_ATTEMPT_HRS} + \\ -0.003002 * \text{SECOND_TERM_EARNED_HRS} + \\ 0.0000007801 * \text{GROSS_FIN_NEED} + \\ -0.0000006846 * \text{COST_OF_ATTEND} + \\ -0.0000008469 * \text{EST_FAM_CONTRIBUTION} + \\ + -39.31$$

Equation For Factor-5

$$0.8357 * \text{STDNT_AGE} + \\ 0.0003496 * \text{FIRST_TERM} + \\ 0.0003496 * \text{SECOND_TERM} + \\ -0.001713 * \text{DISTANCE_FROM_HOME} + \\ 0.5833 * \text{HIGH_SCHL_GPA} + \\ 0.02436 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ 0.06831 * \text{FIRST_TERM_EARNED_HRS} + \\ -0.1323 * \text{SECOND_TERM_ATTEMPT_HRS} + \\ -0.02792 * \text{SECOND_TERM_EARNED_HRS} + \\ -0.0000005283 * \text{GROSS_FIN_NEED} + \\ 0.0000001134 * \text{COST_OF_ATTEND} + \\ 0.0000006384 * \text{EST_FAM_CONTRIBUTION} + \\ + -156.2$$

Equation For Factor-6

$$1.051 * \text{STDNT_AGE} + \\ -0.0001415 * \text{FIRST_TERM} + \\ -0.0001415 * \text{SECOND_TERM} + \\ 0.001722 * \text{DISTANCE_FROM_HOME} + \\ -1.078 * \text{HIGH_SCHL_GPA} + \\ 0.2453 * \text{FIRST_TERM_ATTEMPT_HRS} + \\ 0.03726 * \text{FIRST_TERM_EARNED_HRS} + \\ -0.01922 * \text{SECOND_TERM_ATTEMPT_HRS} +$$

$$\begin{aligned}
& -0.01892 * \text{SECOND_TERM_EARNED_HRS} + \\
& 0.00000002823 * \text{GROSS_FIN_NEED} + \\
& -0.00000007311 * \text{COST_OF_ATTEND} + \\
& -0.00000008349 * \text{EST_FAM_CONTRIBUTION} + \\
& + 37.85
\end{aligned}$$

Equation For Factor-7

$$\begin{aligned}
& -0.5164 * \text{STDNT_AGE} + \\
& 0.0001226 * \text{FIRST_TERM} + \\
& 0.0001226 * \text{SECOND_TERM} + \\
& 0.002142 * \text{DISTANCE_FROM_HOME} + \\
& 0.8997 * \text{HIGH_SCHL_GPA} + \\
& 0.1157 * \text{FIRST_TERM_ATTEMPT_HRS} + \\
& 0.1002 * \text{FIRST_TERM_EARNED_HRS} + \\
& -0.2703 * \text{SECOND_TERM_ATTEMPT_HRS} + \\
& -0.09426 * \text{SECOND_TERM_EARNED_HRS} + \\
& -0.00000009062 * \text{GROSS_FIN_NEED} + \\
& -0.00000003567 * \text{COST_OF_ATTEND} + \\
& -0.00000002316 * \text{EST_FAM_CONTRIBUTION} + \\
& + -40.84
\end{aligned}$$

Equation For Factor-8

$$\begin{aligned}
& 0.8309 * \text{STDNT_AGE} + \\
& -0.00004512 * \text{FIRST_TERM} + \\
& -0.00004512 * \text{SECOND_TERM} + \\
& 0.001948 * \text{DISTANCE_FROM_HOME} + \\
& 1.011 * \text{HIGH_SCHL_GPA} + \\
& -0.3883 * \text{FIRST_TERM_ATTEMPT_HRS} + \\
& -0.04836 * \text{FIRST_TERM_EARNED_HRS} + \\
& 0.04474 * \text{SECOND_TERM_ATTEMPT_HRS} + \\
& 0.06918 * \text{SECOND_TERM_EARNED_HRS} + \\
& 0.000000112 * \text{GROSS_FIN_NEED} + \\
& 0.00000004583 * \text{COST_OF_ATTEND} + \\
& -0.0000000749 * \text{EST_FAM_CONTRIBUTION} + \\
& + 4.255
\end{aligned}$$

Equation For Factor-9

$$\begin{aligned}
& 0.2762 * \text{STDNT_AGE} + \\
& -0.0004363 * \text{FIRST_TERM} + \\
& -0.0004363 * \text{SECOND_TERM} + \\
& 0.0001137 * \text{DISTANCE_FROM_HOME} + \\
& 1.386 * \text{HIGH_SCHL_GPA} + \\
& 0.4684 * \text{FIRST_TERM_ATTEMPT_HRS} +
\end{aligned}$$

$$\begin{aligned}
& -0.2976 * \text{FIRST_TERM_EARNED_HRS} + \\
& 0.1615 * \text{SECOND_TERM_ATTEMPT_HRS} + \\
& -0.1001 * \text{SECOND_TERM_EARNED_HRS} + \\
& 0.00000007674 * \text{GROSS_FIN_NEED} + \\
& -0.00000006234 * \text{COST_OF_ATTEND} + \\
& 0.00000002993 * \text{EST_FAM_CONTRIBUTION} + \\
& + 161.8
\end{aligned}$$

The total Variance Explained:

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.301	19.172	19.172	2.301	19.172	19.172
2	2.083	17.356	36.528	2.083	17.356	36.528
3	1.746	14.548	51.076	1.746	14.548	51.076
4	1.087	9.055	60.131	1.087	9.055	60.131
5	1.060	8.837	68.968	1.060	8.837	68.968
6	1.030	8.585	77.553	1.030	8.585	77.553
7	.911	7.592	85.145	.911	7.592	85.145
8	.865	7.204	92.350	.865	7.204	92.350
9	.477	3.978	96.327	.477	3.978	96.327
10	.358	2.985	99.312			
11	.083	.688	100.000			
12	-1.110E-16	-9.252E-16	100.000			

Extraction Method: Principal Component Analysis.

Table 5. Using the PCA Methods for 12 Variables

While using the PCA methods for the following 12 variables: STDNT_AGE, FIRST_TERM, SECOND_TERM, DISTANCE_FROM_HOME, HIGH_SCHL_GPA, FIRST_TERM_ATTEMPT_HRS, FIRST_TERM_EARNED_HRS, SECOND_TERM_ATTEMPT_HRS, SECOND_TERM_EARNED_HRS, GROSS_FIN_NEED COST_OF_ATTEND and EST_FAM_CONTRIBUTION.

The above total variance was extracted that explains for example if we use 4 variables instead of the 12 variables we will get 60.13% variance. Also, if we use 8 variables for instance, it would give us 92.35% variance and ect.

4.2.4 Data Visualization

The dataset distribution for some fields are presented below:

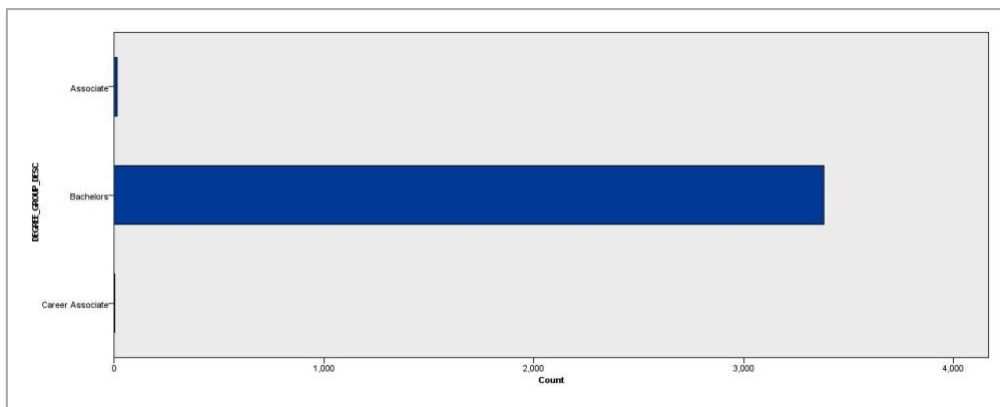


Figure 3. The Distribution of DEGREE_GROUP_DESC

The above graph illustrates the The Distribution of DEGREE_GROUP_DESC for the values (Associate, Bachelors,, and Career Associate). As the most proportion count goes to the Bachelor's students that is our main concern for this study which includes a total of 99.53%. However, if we notice the Associate proportion holds a total of 0.35% and the lowest proportion goes to Career Associate which counts for 0.12% only.

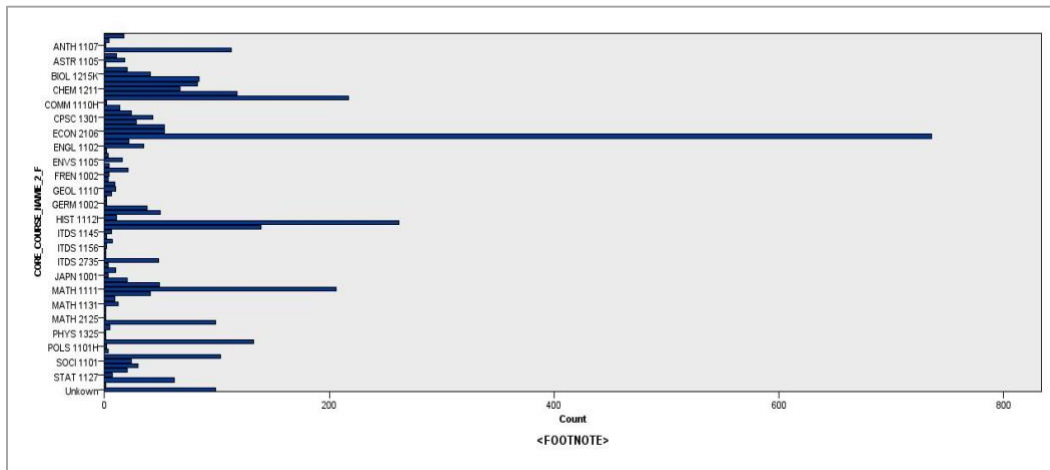


Figure 4. Distribution of CORE_COURSE_NAME_2_F

In this dataset, its observable that some courses are mostly studied by students at the university such as, English 1101 is a total of 736 used compared to other courses for example, PSYC a total count of 103 only. Also, history course is the second big course that is preferred by students and a count of 262 most likely to this course. The data distribution of the courses that are preferred by students are illustrated in the above graph.

The general overview of the dataset values somehow needs some effort to prepare the data for proper analysis and use. The current dataset variables below include some missing values.

(Student_Age) includes missing values for the age group between (20-26) years old. The only numbers are mentioned in the dataset is for the age of 17, a total count of (309) and age 18 includes (2860) counts and for the age 19 is only mentioned (190) count. The remaining age group data is missing.

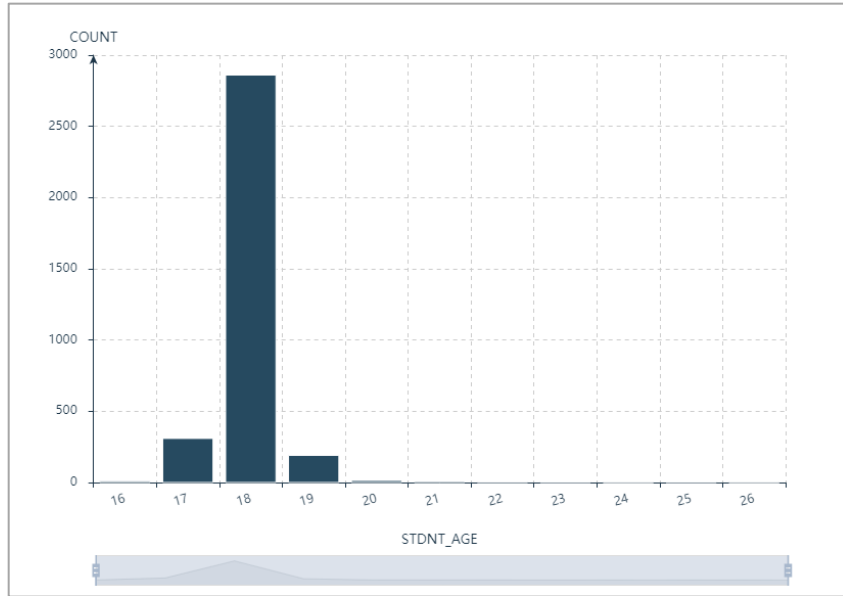


Figure 5. (STDNT_AGE) Variable Missing Data

(DISTANCE_FROM_HOME) variable include null blank values, as the below figure shows that values for the long distance from home are not included.

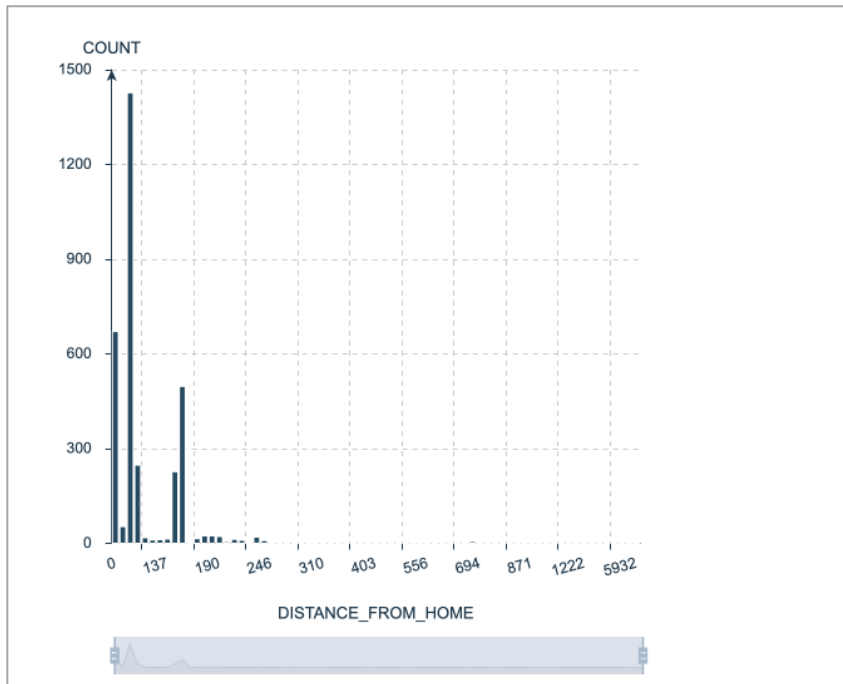


Figure 6. (DISTANCE_FROM_HOME) Variable Missing Data

(HIGH_SCHL_GPA) variable have some null and missing values as the histogram shows that missing value in the field of high school gpa is not mentioned and the distribution of the data is between values (2-4), values between (0.10 and 1) are missing.

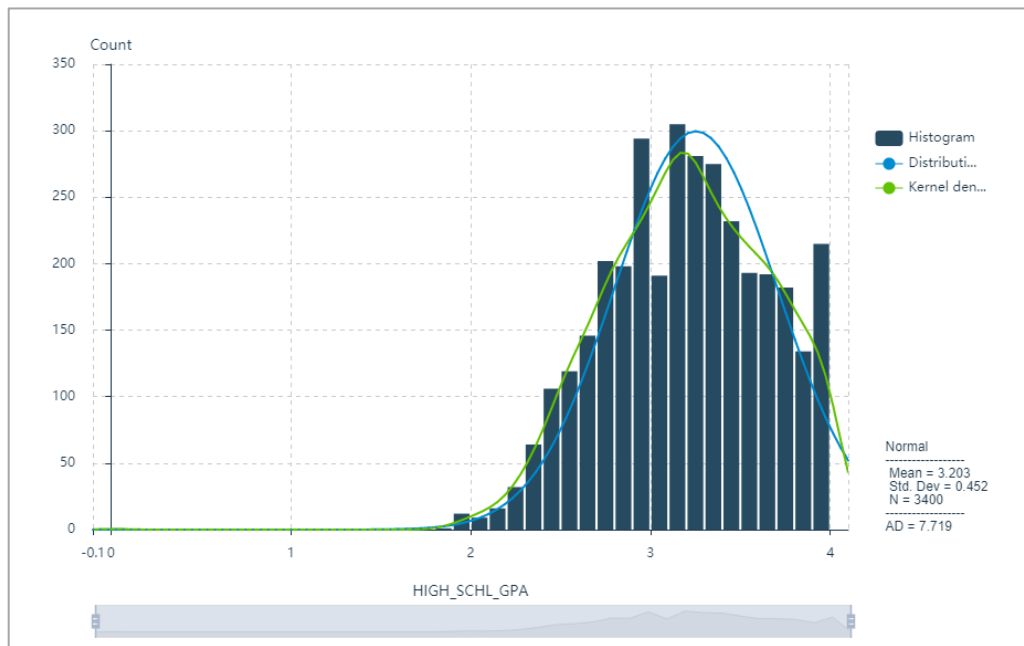


Figure 7. (HIGH_SCHL_GPA) Variable Missing Data

These were some real example of the dataset challenges that requires some attention prior data analysis, the methods to address these issues were employed to ensure data quality. Replacement of values and using machine learning techniques to address missing values were used to enhance our dataset.

The following variables were handled as mentioned above:

(Student_Age) a grouping of three main categories were fixed for students age and included main three groups reclassification.

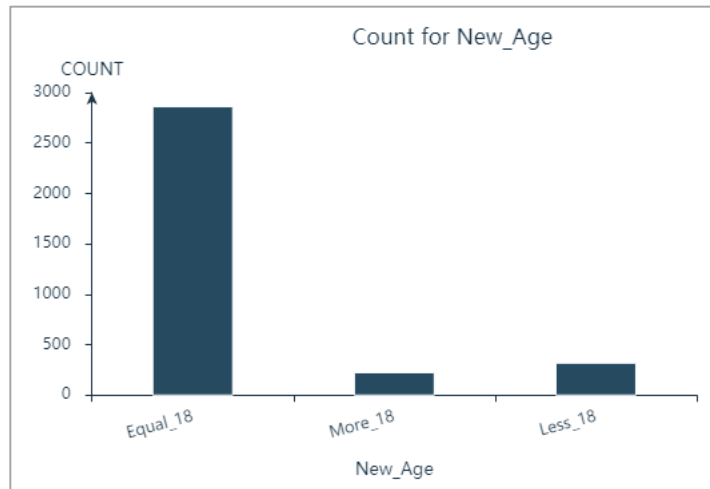


Figure 8. The New Group Variable Named as (New_Age)

For student age, we came up with a new age group range between (16-18), (more than 18) and less than 18 years). Also, the student's degree studying Bachelor is only selected, since this is our main focus of the study.

	STUDENT IDENTIFIER	STDNT_AGE	STDNT_GENDER	STDNT_BACKGROUND
1	7808615.000	18.000	F	BGD 1
2	7830063.000	19.000	F	BGD 1
3	7847538.000	18.000	M	BGD 1
4	8006429.000	18.000	M	BGD 1
5	7962680.000	18.000	F	BGD 1
6	7815697.000	18.000	M	BGD 1
7	7838856.000	18.000	F	BGD 1
8	7960448.000	18.000	F	BGD 1
9	7944779.000	18.000	F	BGD 3
10	7902044.000	18.000	F	BGD 1

Table 6. Student New Age Group Classification

(DISTANCE_FROM_HOME) In our case we have used the Model creation for handling the missing values, we used machine learning technique to replace the missing values through CRT.

This section shows the relationship among the variables in the dataset, as some variables are correlated with each other, and some are not. The variables that have impact on the Return of the students to the complete their studies.

Field	1.000*	0.000*	Importance
STDNT_AGE	18.015	17.986	0.798 Unimportant
FIRST_TERM	200787.583	200774.216	0.939 Marginal
SECOND_TERM	200881.583	200868.216	0.939 Marginal
DISTANCE_FROM_HOME	111.604	96.530	0.874 Unimportant
HIGH_SCHL_GPA	3.135	3.222	1.000 Important
FIRST_TERM_ATTEMPT_HRS	13.996	13.989	0.082 Unimportant
FIRST_TERM_EARNED_HRS	11.821	12.313	1.000 Important
SECOND_TERM_ATTEMPT_HRS	13.789	14.354	1.000 Important
SECOND_TERM_EARNED_HRS	11.807	12.709	1.000 Important
GROSS_FIN_NEED	308523.308	302569.279	0.230 Unimportant
COST_OF_ATTEND	549101.000	553222.523	0.128 Unimportant
EST_FAM_CONTRIBUTION	313880.417	353690.225	0.771 Unimportant
UNMET_NEED	91667.125	66868.411	0.965 Important

Table 7. Shows The Importance of Each Variable That Have Impact on the Return of The Students

As per the above table which shows the importance of each variable that have impact on the Return of the students to the complete their studies. In which 5 variables are considered to be an important factor that are HIGH_SCHL_GPA, UNMET_NEED, SECOND_TERM_EARNED_HRS, SECOND_TERM_ATTEMPT_HRS and FIRST_TERM_EARNED_HRS.

The below Boxplot was created to measure the importance of each variable that have impact on the Return of the students in which seems to take on two values (1 and 0, likely representing 'Yes' and 'No') 1 means returned and 0 means did not return to the complete their studies.

The mentioned variables are HIGH_SCHL_GPA, UNMET_NEED, SECOND_TERM_EARNED_HRS, SECOND_TERM_ATTEMPT_HRS and FIRST_TERM_EARNED_HRS against a binary outcome (RETURNED_2ND_YR).

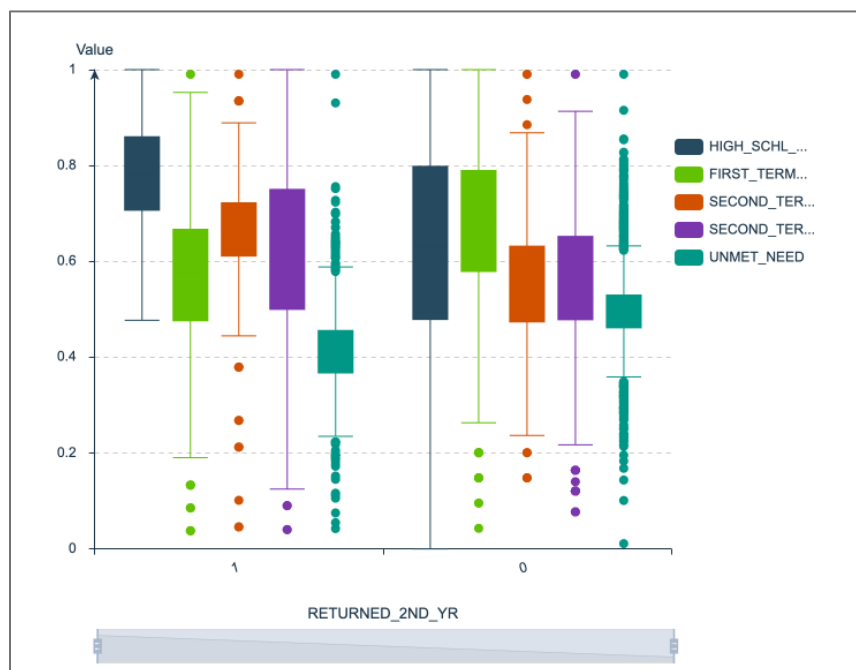


Figure 9. Boxplot of Variables that Have Impacted on the Return of The Students

Here's a breakdown of how to interpret the box plot:

1. **Box Plot Components:** Each colored box represents a category. The bottom and top of a box are the first (Q1) and third quartiles (Q3), respectively, and the band inside the box is the median. The "whiskers" extend from the box to the smallest and largest values, excluding outliers. Outliers are

plotted as individual points which are thought to originate via an entirely distinct distribution over the rest of data set, may seriously affect statistical analysis as well as frequently result in incorrect conclusions (Schwertman, N. C., Owens, M. A., & Adnan, R. 2004).

2. Category Analysis:

HIGH_SCHL_GPA: There's a significant difference when comparing the 1 (Yes) and 0 (No) outcomes. The median for students who returned for the second year is higher than those who did not.

FIRST_TERM_EARNED_HRS: Similar to HIGH_SCHL..., the medians are higher for those who returned, but the spread (interquartile range) is narrower, indicating less variability among students who returned.

SECOND_TERM_EARNED_HRS: The median for those who returned is slightly higher than for those who did not, with a wider spread for the former.

3. Comparative Analysis:

The spread of values (IQR) for each category varies, which can indicate differing levels of variability between these categories.

Outliers are present in each category, indicating that there are values that fall significantly outside the typical range. This could suggest exceptional cases or data entry errors.

4. Outcome Analysis:

The binary outcome (RETURNED_2ND_YR) shows a clear trend where the median values for those who returned (1) are generally higher than those who did not (0), across all categories.

5. Statistical Significance: Without additional context, we can't determine if the differences observed are statistically significant.

This plot might be used to assess factors affecting whether students return for a second year of study, with the categories representing different metrics or characteristics measured. The higher values for students who returned suggest that these metrics might be positively correlated with the likelihood of returning.

4.3 Machine Learning Model Development:

4.3.1 Feature Selection for Building Models

We have used the feature selection technique which is the process of choosing a subset of relevant features (variables, attributes) from the original set of features in a dataset. The objective of feature selection is to improve the performance of machine learning models by reducing the dimensionality of the data while retaining the most informative and discriminative features and to support in putting the best model.

	Rank #	Field	Measurement	Importance	Value
<input checked="" type="checkbox"/>	1	A CORE COURSE GRADE 1 S	Nominal	Important	1.0
<input checked="" type="checkbox"/>	2	A CORE COURSE NAME 1 S	Nominal	Important	1.0
<input checked="" type="checkbox"/>	3	A CORE COURSE GRADE 2 S	Nominal	Important	1.0
<input checked="" type="checkbox"/>	4	A CORE COURSE GRADE 1 F	Nominal	Important	1.0
<input checked="" type="checkbox"/>	5	A CORE COURSE GRADE 2 F	Nominal	Important	1.0
<input checked="" type="checkbox"/>	6	A CORE COURSE GRADE 3 F	Nominal	Important	1.0
<input checked="" type="checkbox"/>	7	SECOND TERM ATTEMPT H...	Continuous	Important	1.0
<input checked="" type="checkbox"/>	8	SECOND TERM EARNED HRS	Continuous	Important	1.0
<input checked="" type="checkbox"/>	9	HIGH SCHL GPA	Continuous	Important	1.0
<input checked="" type="checkbox"/>	10	DISTANCE FROM HOME	Ordinal	Important	1.0
<input checked="" type="checkbox"/>	11	FIRST TERM EARNED HRS	Continuous	Important	1.0
<input checked="" type="checkbox"/>	12	A STDNT MAJOR	Nominal	Important	0.998
<input checked="" type="checkbox"/>	13	A HOUSING STS	Flag	Important	0.994
<input checked="" type="checkbox"/>	14	A STDNT BACKGROUND	Nominal	Important	0.992
<input checked="" type="checkbox"/>	15	A New Age	Ordinal	Important	0.977
<input checked="" type="checkbox"/>	16	STDNT AGE	Ordinal	Important	0.976
<input checked="" type="checkbox"/>	17	A CORE COURSE NAME 3 F	Nominal	Important	0.961
<input checked="" type="checkbox"/>	18	FIRST TERM	Ordinal	Important	0.959
<input checked="" type="checkbox"/>	19	SECOND TERM	Ordinal	Important	0.959
<input checked="" type="checkbox"/>	20	UNMET NEED	Continuous	Important	0.955
<input type="checkbox"/>	21	A MOTHER HI EDU DESC	Nominal	Marginal	0.919
<input type="checkbox"/>	22	A STDNT GENDER	Flag	Marginal	0.916
<input type="checkbox"/>	23	A FATHER HI EDU DESC	Nominal	Unimportant	0.776
<input type="checkbox"/>	24	EST FAM CONTRIBUTION	Continuous	Unimportant	0.743
<input type="checkbox"/>	25	GROSS FIN NEED	Continuous	Unimportant	0.19
<input type="checkbox"/>	26	A CORE COURSE NAME 2 F	Nominal	Unimportant	0.133
<input type="checkbox"/>	27	COST OF ATTEND	Continuous	Unimportant	0.099
<input type="checkbox"/>	28	FIRST TERM ATTEMPT HRS	Continuous	Unimportant	0.047
<input type="checkbox"/>	29	A CORE COURSE NAME 1 F	Nominal	Unimportant	0.03

Table 8. Showing The 22 Important Variables That are Used in Order to Build Up he Model

In order to build up the model, we will use the important above 22 fields.

Field	Measurement	Values	Missing	Check	Role #
FIRST TERM	Ordinal	200508.0,200...		None	Input
A CORE COURSE ...	Nominal	"" ,ANTH 1105...		None	Input
A CORE COURSE ...	Nominal	"" ,A,B,C,D,F,IN...		None	Input
A CORE COURSE ...	Nominal	"" ,ANTH 1105...		None	Input
A CORE COURSE ...	Nominal	"" ,A,B,C,D,F,IN...		None	Input
A CORE COURSE ...	Nominal	"" ,ANTH 1105...		None	Input
A CORE COURSE ...	Nominal	"" ,A,B,C,D,F,IN...		None	Input
SECOND TERM	Ordinal	200602.0,200...		None	Input
A CORE COURSE ...	Nominal	"" ,ANTH 1105...		None	Input
A CORE COURSE ...	Nominal	"" ,A,B,C,D,F,IN...		None	Input
A CORE COURSE ...	Nominal	"" ,A,B,C,D,F,IN...		None	Input
HIGH SCHL GPA	Continuous	[0.0,4.0]		None	Input
FIRST TERM EA...	Continuous	[0.0,21.0]		None	Input
SECOND TERM ...	Continuous	[2.0,23.0]		None	Input
SECOND TERM ...	Continuous	[0.0,23.0]		None	Input
\$F-Factor-1	Continuous	[-4.64549653...		None	Input
\$F-Factor-2	Continuous	[-3.70146470...		None	Input
\$F-Factor-3	Continuous	[-2.56775008...		None	Input
\$F-Factor-4	Continuous	[-1.64937623...		None	Input
\$F-Factor-5	Continuous	[-3.03040712...		None	Input
\$F-Factor-6	Continuous	[-13.9491730...		None	Input
RETURNED 2ND...	Flag	1.0/0.0		None	Target

Table 9. List of Variables Used for Scenario Three

4.3.2 Machine Learning Models Used

In this thesis we have used LVSM 1, VSM 1, Random Tree, Logistic Regression and Neural Network. Logistic regression was chosen due to its simplicity, interpretability, and effectiveness in binary classification tasks. LSVM was selected for its ability to handle linearly separable data and its effectiveness in classifying instances into two classes. Neural networks were selected for their ability to capture complex, nonlinear relationships in the data. Random trees were considered for their simplicity, interpretability, and ability to capture non-linear relationships in the data.

4.3.3 Validation And Testing Procedures

4.3.3.1 Partitioning

In the practice of data partitioning, we partitioned the data into two datasets, the training dataset and the testing dataset. The training dataset encapsulates 70% of the total dataset, whereas the testing dataset embodies the remaining 30%. The next step was that we have build a model dataset to train and evaluate predictive models. The process involves selecting relevant input variables that are expected to have an impact on the target variable, which is the outcome we aim to predict. By using historical data where both input variables and the target variable are known, we can train a model to learn the relationship between the inputs and the target. Once trained, the model can then be used to make predictions. In order to build the model, we have selected 35 variables and 3,384 cases. The training dataset includes 2,310 cases and the testing dataset includes 1,074 cases.

4.3.3.2 Balancing

In this section there will be a balancing for the data which will be the process of adjusting the distribution of classes or categories within a dataset to mitigate biases and improve the performance of machine learning models. When a dataset is imbalanced, meaning that one class is significantly more prevalent than others, it can lead to issues such as biased predictions, poor generalization, and lower accuracy on minority classes (Mduma, N. 2023).

However, this is the case for our data since it's not balanced in which we have 0 more than 1 as it is showing in the below capture. In order to avoid the un-balanced data, we have used SMOTE technique to balance the dataset. SMOTE is a popular method used to address class imbalance in machine learning datasets, particularly in binary classification problems where one class is significantly underrepresented compared to the other (Dablain, D., Krawczyk, B., & Chawla, N. V. 2022). SMOTE works by generating synthetic samples for the minority class to balance the class distribution.

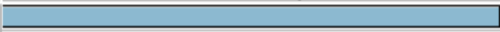

Table Graph Annotations			
Value ✓	Proportion	%	Count
0.000		78.72	2664
1.000		21.28	720

Figure 10. Showing the Proportion of the Student's Return Before the SMOTE Technique

As shown below in the table, we have oversampled the RETURNED_2ND_YR = 1, by to 3.6 in order to have a balanced partitioned dataset.

Balance Directives:	
Factor	Condition
1.0	RETURNED_2ND_YR = 0.0
3.6	RETURNED_2ND_YR = 1.0

Figure 11. Showing Balance Directives Factors of Student's Return During SMOTE Technique

As a result, the proportion of the value 1 and value zero, in the training dataset of our target RETURNED_2ND_YR, has been balanced to a distribution of 50.78% and 49.22%.



Value	Proportion	%	Count
0.000		49.22	1791
1.000		50.78	1848

Figure 12. Showing the Proportion of the Student's Return After the SMOTE Technique

4.3.3.3 Evaluation Metric to Assess Model Performance

4.3.3.3.1 Confusion Matrix

Within machine learning as well as statistics, there is confusion matrix called error matrix that is a technique used to assess how well categorization models perform. This allows for a thorough examination of classifier's performance throughout several classes by summarizing the predictions generated by classifier in comparison to actual group labels.

		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

Figure 13. Confusion Matrix Explaining the Predicted and the Actual

- TP: True Positives - The number of patients with the disease correctly predicted as "yes."
- TN: True Negatives - The number of patients without the disease was correctly predicted as "no."
- FP: False Positives - The number of patients who don't have the disease but were incorrectly predicted as "yes."
- FN: False Negatives - The number of patients who have the disease but were incorrectly predicted as "no."

4.3.3.3.2 Accuracy

One of the most important evaluation metrics for determining how well classification model performs generally is accuracy. This can be expressed as the proportion of accurately predicted instances to all occurrences in dataset. The formula accuracy calculation is:

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

4.3.3.3.3 Precision and Recall

In order to comprehend the trade-off among false positives as well as false negatives within machine learning, precision along with recall are crucial assessment measures. The below formulas explain precision and recall:

$$P = \text{Precision} = \frac{TP}{(TP + FP)}$$

$$R = \text{Recall} = \frac{TP}{(TP + FN)}$$

Precision (P): The percentage of true-positive predictions throughout all positive forecasts is known as precision (P). It is a gauge of the degree of accuracy of the optimistic forecasts.

Recall (R): The percentage of genuine positive predictions throughout all real positive cases is called recall (R), which is often referred to as sensitivity or true-positive rate (TPR). It assesses how well the classifier can recognize positive examples.

4.3.3.3.4 F1-score

The F1-score is a metric that balances recall as well as accuracy, calculated as harmonic mean of these two. While working with unbalanced datasets—those in which one class is noticeably more prevalent than the other—it is advantageous. The F1 score is calculated as follows:

$$\text{F1 Score} = \frac{2 * P * R}{(P + R)}$$

The F1-score is a statistic that takes into account false positives as well as false negatives, calculated as harmonic mean of recall as well as accuracy.

4.3.3.3.5 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

One common assessment metric concerning binary classification issues is AUC-ROC. It gauges how well the model can discriminate between positive as well as negative classes. Plotting rate of true positives (recall) versus rate of false positives (1 - specificity) at different categorization thresholds is known as ROC curve. In which the area underneath ROC curve is represented by AUC-ROC, wherein a greater value denotes better model efficiency and performance.

4.3.4 Results

4.3.4.1 Presentation of Experimental Results

While building the models, we have utilized diverse scenarios to choose the input for the variables for several reasons. Primarily, it helps account for the inherent uncertainty and variability present in real-world data. By considering various scenarios, we can capture a broader range of possible outcomes and ensure our models are robust and adaptable. Additionally, exploring different scenarios (as shown in the Appendix) allows us to assess the sensitivity of the model to different inputs and assumptions, thereby enhancing its reliability and effectiveness in decision-making. Lastly, incorporating multiple scenarios enables us to anticipate and prepare for potential future changes or disruptions, thereby mitigating risks and improving overall model performance.

In this scenario as demonstrated in the below figure and table, using the feature selection method with 21 variables 5 models have been used. The overall accuracy for the SVM 1 is % is 90.928 and the AUC is 0.925 with 21 fields resulted to be the best model in this scenario.

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	No. Fields Used	Overall Accuracy (%)	Accumulated Accuracy (%)	Area Under Curve	Accumulated AUC	Precision	Recall
<input checked="" type="checkbox"/>		SVM 1	< 1	0.0	0	2.963	21	90.928	90.928	0.925	0.925	0.921	0.628
<input checked="" type="checkbox"/>		Random Trees 1	< 1	0.0	0	2.629	21	86.318	86.318	0.885	0.885	0.712	0.6
<input checked="" type="checkbox"/>		Logistic regressi...	< 1	0.0	0	2.25	21	84.043	84.043	0.821	0.821	0.746	0.379
<input checked="" type="checkbox"/>		LSVM 1	< 1	0.0	0	2.246	21	83.983	83.983	0.817	0.817	0.762	0.360
<input type="checkbox"/>		Neural Net 1	< 1	0.0	0	1.945	21	82.004	82.004	0.735	0.735	0.745	0.270

Table 10. List of Models used for Scenario Three

4.3.4.2 Evaluation of Predictor Importance

Notation

The following notation is used throughout this chapter unless otherwise stated:

Y Target

X_j Predictor, where $j=1, \dots, k$

K The number of predictors

$Y = f(X_1, X_2, \dots, X_k)$ Model for Y based on predictors X, through X_k

Variance Based Method

Predictors are ranked according to the sensitivity measure defined as follows:

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

Where $V(Y)$ is the unconditional output variance. In the numerator, the expectation operator E calls for an integral over X_i that is, over all factors but X , then the variance operator V implies a further integral over X .

Predictor importance is then computed as the normalized sensitivity.

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

4.3.4.3 Predictor Importance of The SVM 1 Model:

This table shows the normalized sensitivity measure calculated by the formula above. For example, HIGH_SCHOOL_GPA with normalized sensitivity that equals 0.09 has a big impact on the attrition. However, CORE_COURSE_GRADE_2_S has sensitivity of 0.06.

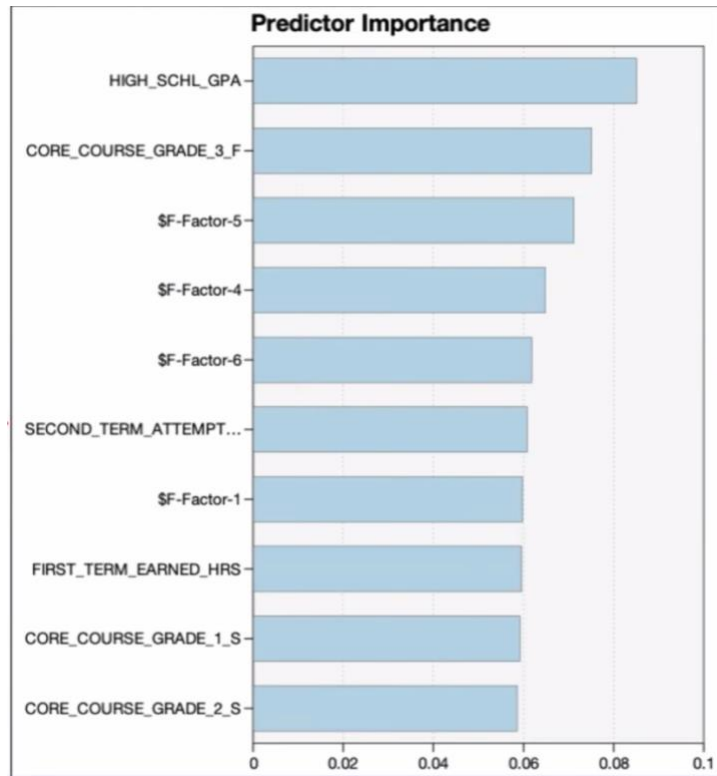


Figure 14. Predictor Importance for SVM 1 in Scenario Three

4.3.4.4 Visualization and Statistical Testing of the Important Predictors

The statistical testing of the numerical variable with the RETURNED_2ND_YR

We will calculate the P-value of the null hypothesis using the Mann-Whitney U Test.

<i>Hypothesis Test Summary</i>				
	Null Hypothesis	Test	Sig. ^{a,b}	Decision
1	The distribution of HIGH_SCHL_GPA is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

2	The distribution of FIRST_TERM_EARNED_HRS is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.
3	The distribution of SECOND_TERM_ATTEMPT_HRS is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.
4	The distribution of Factor1 is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.
5	The distribution of Factor4 is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	.158	Retain the null hypothesis.
6	The distribution of Factor5 is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	.664	Retain the null hypothesis.
7	The distribution of Factor6 is the same across categories of Class_Predicted_RETURNED_2ND_YR.	Independent-Samples Mann-Whitney U Test	<.001	Reject the null hypothesis.

Table 11. Hypothesis Test Summary Using Mann-Whitney

Below the distribution of the variable where the P-value < 0.001 (Rejected the null hypothesis) which are 5.

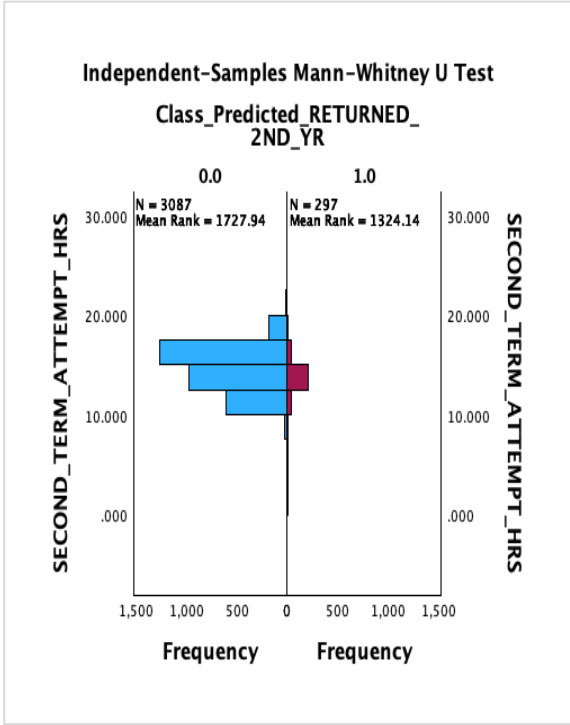


Figure 15. The Distributions of the Variable SECOND_TERM_ATTEMPT_HRS

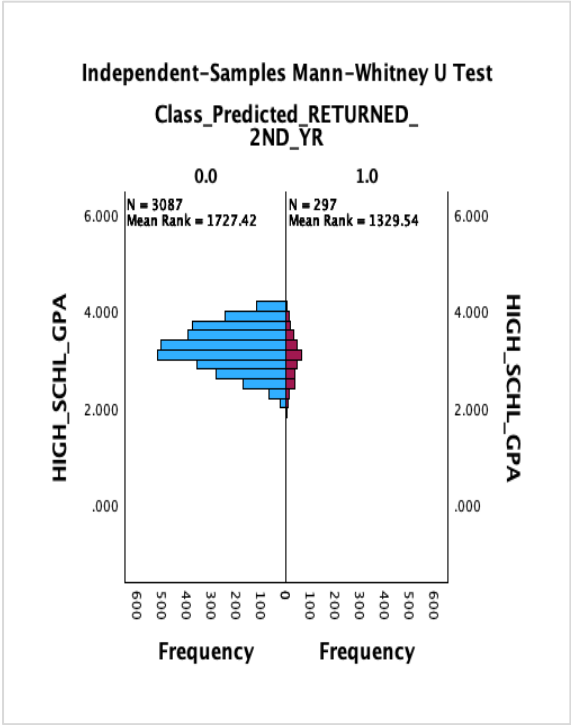


Figure 16. The Distributions of the Variable HIGH_SCHL_GPA

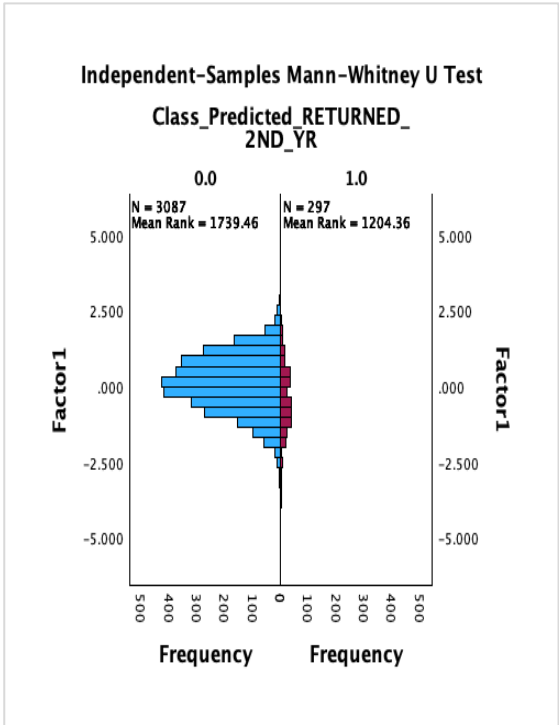


Figure 18. The Distributions of the Variable Factor 1

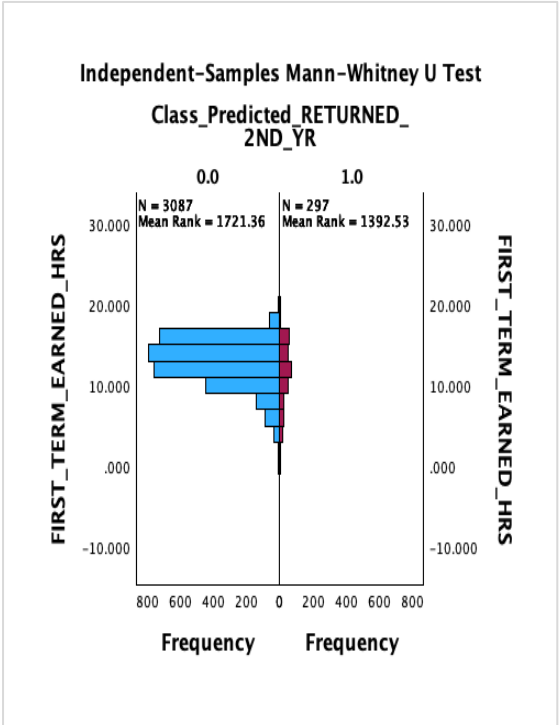


Figure 17. The Distributions of the Variable FIRST_TERM_EARNED_HRS

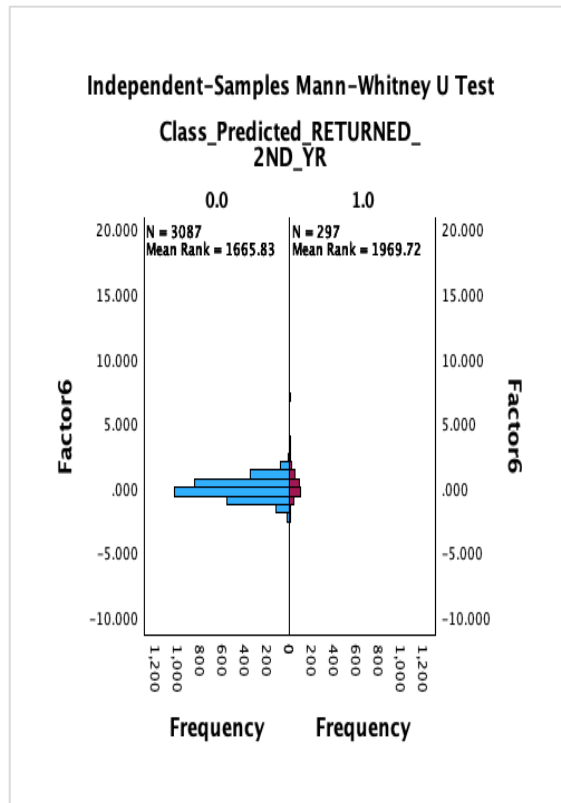


Figure 19. The Distributions of the Variable Factor 6

The distributions of the variables are not the same in the RETURNED_2ND_YR.

The statistical testing of the nominal variable with the RETURNED_2ND_YR.

We will calculate the P-value of the null hypothesis using the Chi-Square.

<i>Chi-Square Tests</i>			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	130.999 ^a	5	<.001
Likelihood Ratio	95.753	5	<.001
N of Valid Cases	3384		

Table 12. P-value of the null hypothesis using the Chi-Square

- a. 1 cells (8.3%) have expected count less than 5. The minimum expected count is 2.46.

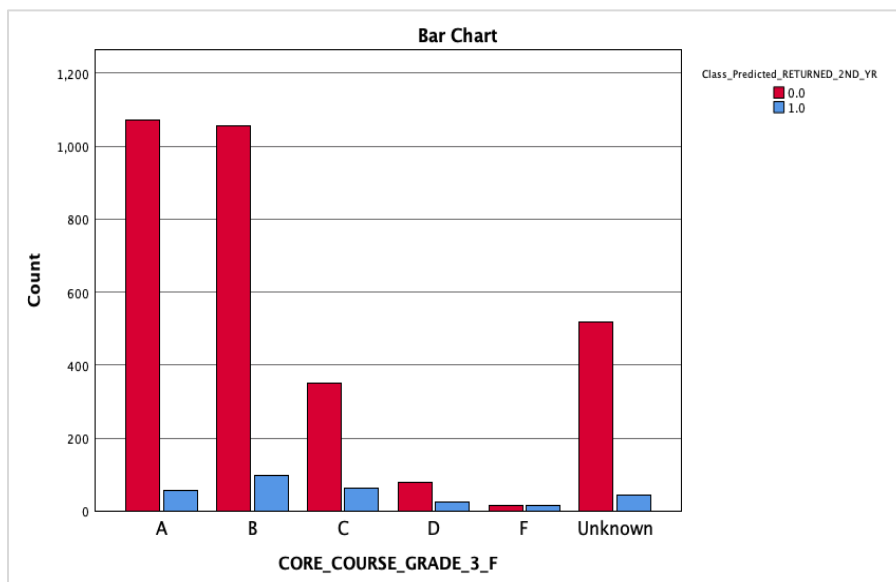


Figure 20. P-Value < 0.001 of The Chi-Square of the Relation Between the CORE_COURSE_GRADE_3_F and the RETURNED_2ND_YR

This chart shows regarding the P-value < 0.001 of the Chi-square leads to high significant relation between the CORE_COURSE_GRADE_3_F and the RETURNED_2ND_YR.

4.3.4.5 Analysis of ROC Curves and AUC Values

When it concerns binary classification roles, the Receiver Operating Characteristic Area Under the Curve (ROC AUC) is essential since it provides important information about the discriminatory capacity models of the logistic regression over various probability levels. Using empirical data indicators via Sensitivity as well as 1-Specificity, we comprehensively examine the idea of ROC AUC along with its use within predictive modeling inside this thesis.

The performance or the results of logistic regression while discrimination threshold varies is graphically represented by the ROC curve. Also in our study the experimental setup upon a threshold of 0.5, wherein the False-Positive Rate (FPR or 1 - Specificity) is contrasted with the True-Positive Rate (TR or Sensitivity). This curve illustrates the fine balance between correctly classifying negatives as well as reliably detecting genuine positives, with each data point representing a different threshold level.

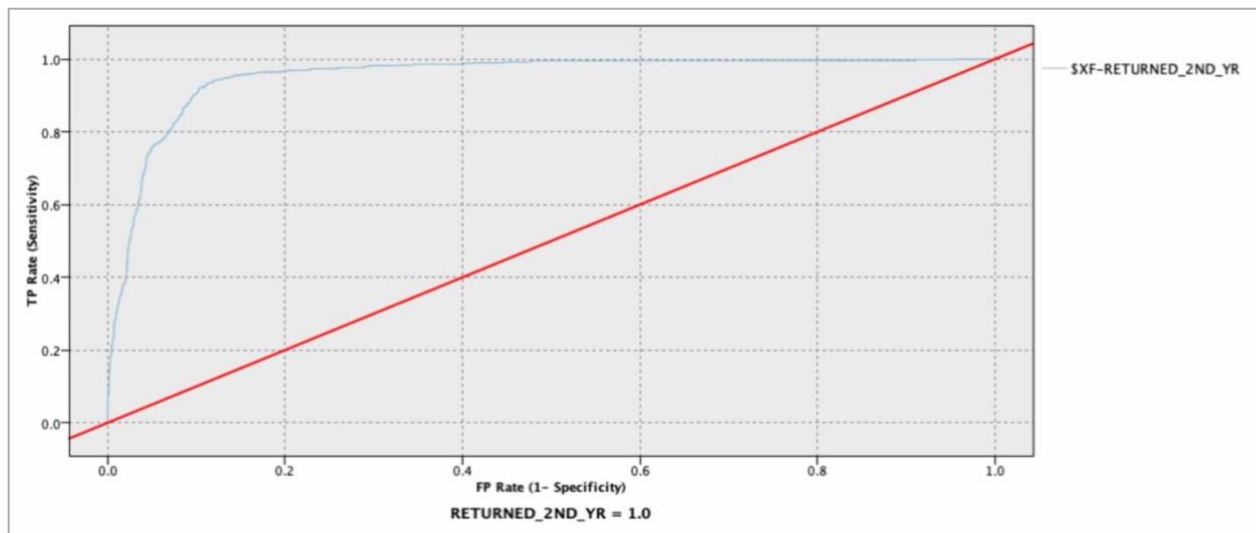


Figure 21. ROC for SVM 1 in Scenario Three

Chapter 5 Discussion

This study aimed to implement an Artificial Intelligence model with the several machines learning techniques assisting the model itself, such as feature selection technique to predict accurately student attrition.

The objectives of this study are:

1. Assess the machine learning models effectiveness to predict student's attrition.
2. Enhance quality of the dataset through data preprocessing techniques and cleaning.
3. Implement feature selection methodology to highlight the main key predictors of student attrition.

The main research question is hitting to assess the performance of different machine learning models in the prediction of student attrition. The outcomes stated that SVM 1 is outstanding other algorithms comparing to each other, having a remarkable accuracy of 90.928 % with precision 0.921 and recall of 0.628. This result illustrated that simpler model could surpass the models with more complexity such as, Random Trees in this context. SVM 1 have high accuracy to implement its effectiveness in identifying students at attrition risk.

With regards to the second research question is centered on improving the dataset quality with the preprocessing and cleaning techniques. The research paper demonstrates an improvement of the dataset quality from 67.65 % to a rapid improvement reaching to 100% score for both cases and features. The result of the achieved score underscores the importance of the data quality in

predictive modeling and data efficiency through data preprocessing and cleaning to ensure accurate results.

Moving to the third research questions is employed to utilize the automated feature selection methodology to pinpoint the most important key predictors of student attrition case. The study engaged with several scenarios to automatically choose fields focused mainly for prediction purpose. This method allowed to generate effective models focusing on relatable features. However, the exact techniques employed for feature selection were not specified, the focus on automated selection methods suggests a methodical approach to identify significant predictors.

Aligning with the literature review in this study, Thammasiri, & Kasap (2014) clarified that for balancing the data, they have oversampled, and then they employed three different classifiers for evaluation as well as prediction. Models contained logistic regression, decision trees, as well as support vector-machines. The findings indicate that when it comes to identifying students who are at threat of attrition, logistic regression and SVM algorithms outperform the decision tree when it comes to classification accuracy.

Our research's emphasis on preprocessing and data cleaning is consistent with previous research emphasizing the critical role that high-quality data plays in modelling predictions (Pintelas, P. E. 2006). Our results, which obtained a flawless data quality score, highlight the need of thorough data preparation methods in ensuring the validity and dependability of prediction models (David, 2010).

In the automated feature selection approaches, the literature supporting methods that expedite the process of building models and give priority to appropriate predictors exists in accordance with our investigation (Guyon and Elisseeff, 2003; Saeys et al., 2007). The use of methodologies for automatically identifying significant features is indicative of a larger machine learning trend that

emphasizes the use of computational methods towards feature selection in order to improve the interpretability as well as generalization of models (Hastie et al., 2009).

Among the Most Significant Features Found Using Predictor Importance: Using predictor significance analysis for every machine learning model in usage, the most crucial features that considerably influenced the prediction of student attrition were found. Due to space restrictions, these attributes are not covered in detail here, but they were very important for the models' ability to predict outcomes. In order to obtain insights into successful retention techniques, future research must focus on studying and interpreting these important factors.

Overall, the study effectively highlights each research question, with the strongest results with the performance of logistic regression in having student attrition prediction. Also, the study contributes in underlining the importance of having data quality enhancement techniques and feature selection methodology in developing precise and dependable predictive models for student attrition.

Chapter 6 Conclusions & Future Work

6.1 Conclusion

However, this study involves creating an AI model that utilizes several machine learning models to predict student attrition. With the testing and analysis approaches, it was recognized that SVM 1 was the most effective algorithm holding a 90.928% accuracy rate. The process is followed with data preprocess and cleaning and feature selection that guaranteed predictive models dependency and reliability.

The study employs the current literature illustrating the effectiveness of machines learning models in solving complex issue of student attrition. Through the highlight of logistic regression performance with the comparison to other algorithms and focusing on the importance of data quality and preprocessing techniques. The study provides meaningful guidance for education policymaker and educators to improve student retention rates.

The study represents a practical implantation of AI for the educational field. By having an AI development model can act as an important resource to highlight students at risk of attrition and enabling the implementation of timely interventions. Employing predictive analytics enables educators to customize the support and resources in an effective way and create a well learning place that have a better support for learners and enhance the education system outputs.

Major financial consequences are caused by student attrition, many universities experience serious difficulties due to student attrition. However, examining the data of the students through algorithms of machine learning contributes in minimizing attrition by identifying critical variables that

influence student attrition. Considering Kaggle dataset, mainly I employed different models of classification in order to forecast student attrition.

With the used data, I have managed to uncover the strong relationship between these variables and the outcome that I am trying to predict. By accurately selecting these variables, I can have more confidence in the reliability of my predictions. Additionally, this shows the importance of using data and analytical techniques to inform decision-making processes, as it can lead to more accurate and effective results.

Even though, that the education field is totally a new field but this new knowledge has helped me to become more knowledgeable and competent in my field, and I am able to apply it to various aspects of my work and personal life. I am grateful for the opportunity to learn and grow in this way, and I am excited to continue expanding my understanding and expertise in education.

Due to confidentiality and data security concerns, the organization did not provide a dataset for analysis, so I had to search by my own. This dataset contained numerous attributes with missing values, which made me hesitant to begin working on it. Out of the attributes included, many appeared to be irrelevant to the current project.

Additionally, the dataset had not been updated in two years, rendering the data potentially outdated and less accurate for analysis. Ideally, having more recent data would have ensured a more precise and informative analysis. To combat these challenges, I meticulously reviewed and cleaned the dataset to ensure the reliability and accuracy of my findings.

6.2 Recommendations

Having a strong understanding of business is crucial because it allows individuals to make informed decisions that greatly impact the success of a company. By possessing knowledge about the intricacies of business operations, one can effectively evaluate risks and choose the right models for various projects. It is not simply about accuracy, but also about understanding the implications of false positives and false negatives.

This deeper level of understanding ensures that the data being used is relevant and valuable for decision-making processes. In essence, having a solid foundation of business understanding is the key to making informed and strategic choices that drive success and growth within an organization.

6.3 Future work

In the near future, I want to work together with colleges to provide real-time data concerning students rather than hunting for an outdated dataset upon accessible data. As a result, universities or higher education having a high rate of attrition are likely to benefit through various methods that reduce student attrition. As this study provided valuable insights, there is an opportunity for future exploration and enhancements. Future research could look into having diverse dataset sources and predictive indicators to enhance model accuracy. Also, longitudinal studies that monitors student's progress over the time could provide more understanding of attrition patterns that could assist in having a strategic plan for student retention.

Additionally, I would like to explore the use of machine learning algorithms to predict student attrition earlier on in their academic journey. By analyzing various factors such as attendance,

grades, and student engagement, we can identify at-risk students and provide targeted interventions to improve their likelihood of success. Engagement with educational experts to validate model's effectiveness and acclimation to the educational practices is crucial. Also, testing the model with real dataset from the educational field to assess its performance and level of validity and possibilities of wider application. Taking into consideration the ethical dynamics with the use of AI in education field should be practicing the fairness, transparency and privacy concerns. The proactive tackle of these ethical matters can assist in building a trustworthy AI application in the education field that ensures its beneficial and responsible.

Furthermore, I am interested in conducting more in-depth research on the root causes of student attrition and developing strategies to address these underlying issues. This could involve working closely with students, faculty, and administrators to create a more supportive and inclusive academic environment that fosters student success.

Overall, my future work aims to continue making a positive impact on the higher education sector by reducing student attrition rates and ensuring that all students have the support they need to thrive academically. The development of an AI model for predicting student attrition showcase a significant step towards enhancing student success and retention in the educational field. By leveraging the machine learning and data analytics, educational representatives can do a proactive process to understand and assist students who are at risk of attrition, intimately contributing to an overall improvement of educational results and societal welfare.

References

- Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *International Journal of Data Mining & Knowledge Management Process*, 5(1), 15.
- Ahmed, S., Al-Mansoori, R., Khan, F., & Hassan, N. (2020). A Longitudinal Study of Student Attrition in Higher Education Institutions in the UAE. *UAE Education Research Journal*, 4(2), 123-145.
- Aljohani, O. (2016). Analyzing the Findings of the Saudi Research on Student Attrition in Higher Education. *International Education Studies*, 9(8), 184-193.
<https://doi.org/10.5539/ies.v9n8p184>
- Ananat, E. O., Gassman-Pines, A., & Gibson-Davis, C. M. (2013). The economic costs of dropping out of high school: Evidence from North Carolina. *The Review of Economics and Statistics*, 95(2), 596-611. doi: 10.1162/REST_a_00252
- Attewell, P., Lavin, D., Domina, T., & Levey, T. (2006). New evidence on college remediation. *Journal of Higher Education*, 77(5), 886-924. <https://doi.org/10.1353/jhe.2006.0030>
- Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.
- Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4), 485-540.
- Braxton, J. M., Hirschy, A. S., & McClendon, S. A. (2004). Understanding and reducing college student departure. *ASHE-ERIC Higher Education Report*, 30(3), 1-128.
- Chen, Y. Y., & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences. *Research in Higher Education*, 51(6), 570-596.
- Dablain, D., Krawczyk, B., & Chawla, N. V. (2022). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*.
- Daffertshofer, A., Lamoth, C. J., Meijer, O. G., & Beek, P. J. (2004). PCA in studying coordination and variability: a tutorial. *Clinical biomechanics*, 19(4), 415-428.

DeBerard, M. S., Spielmans, G. I., & Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College Student Journal*, 38(1), 66-80.

Doe, J. (2019). The impact of technology on education. *Educational Psychology*, 45(2), 123-135.

Garrison, D. R. (2017). Online and distance education in the time of change: A study of perceptions of academics. *The International Review of Research in Open and Distributed Learning*, 18(3), 1-18. <https://doi.org/10.19173/irrodl.v18i3.3005>

Grau-Valldosera, J., & Minguillón, J. (2014). Rethinking dropout in online higher education: The case of the Universitat Oberta de Catalunya. *International Review of Research in Open and Distributed Learning*, 15. <https://doi.org/10.19173/irrodl.v15i1.1628>

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3, 1157-1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

https://www.researchgate.net/publication/243103404_Sensitivity_Analysis_in_Practice_A_Guide_to_Assessing_Scientific_Models

Jayawant N. Mandrekar (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, Volume 5, Issue 9.

Jian Chen, Thanh G. Phan, David C. (2010). Ridge penalized logistical and ordinal partial least squares regression for predicting stroke deficit from infarct topography. *Journal of Biomedical Science and Engineering* Vol.3 No.6, June 25, 2010

Johnson, N. (2012). *The Institutional Costs of Student Attrition*. Research Paper. Delta Cost Project at American Institutes for Research. Retrieved from: <https://eric.ed.gov/?id=ED536126>

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International journal of computer science*, 1(2), 111-117.

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15), 3093. <https://doi.org/10.3390/app9153093>

- Maher, M., & Macallister, H. (2013). Retention and attrition of students in higher education: Challenges in modern times to what works. *Challenges in Modern Times to What Works*, 3(2), 62-73. <http://doi.org/10.5539/hes.v3n2p62>
- Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data*, 8(3), 49.
- Pascarella, E. T., & Terenzini, P. T. (1980). Predicting voluntary freshman year persistence/withdrawal behavior in a residential university: A path analytic validation of Tinto's model. *Journal of Educational Psychology*, 72(6), 856-862.
- Persson, P., & Rossin-Slater, M. (2018). The long-term effects of youth joblessness on health: Evidence from Swedish school-leavers. *American Economic Journal: Applied Economics*, 10(3), 171-199. doi: 10.1257/app.20160216
- Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.
- Saltelli, A., S. Tarantola, E. F. Campolongo, and M. Ratto. 2004. *Sensitivity Analysis in Practice - A Guide to Assessing Scientific Models*: John Wiley.
- Saltelli, A. 2002. *Making best use of model evaluations to compute sensitivity indices*. *Computer Physics Communications*, 145:2, 280-297.
- Schwertman, N. C., Owens, M. A., & Adnan, R. (2004). A simple more general boxplot method for identifying outliers. *Computational statistics & data analysis*, 47(1), 165-174.
- Seidman, A. (2005). *College student retention: Formula for student success*. Greenwood Publishing Group.
- Seltzer, R. (2021, September 21). College tuition and fees continue to rise. Inside Higher Ed. <https://www.insidehighered.com/news/2021/09/21/college-tuition-and-fees-continue-rise>
- Shaw, M., Burrus, S. W. M., & Ferguson, K. (2016). Factors that influence student attrition in online courses. *Online Journal of Distance Learning Administration*. https://www.researchgate.net/publication/308310140_Factors_that_Influence_Student_Attrition_in_Online_Courses
- Statista. (2022). Pupils out of lower secondary school by gender and region Brahms Kontor, Hamburg. Retrieved from <https://www.statista.com>

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

Winters, J. V., Xu, W., & Guo, Z. (2019). Estimating the impact of student attrition on university revenues: Evidence from a major public research university. *Research in Higher Education*, 60(5), 542-566. <https://doi.org/10.1007/s11162-018-9537-2>

Zhang, L., & Johnes, G. (2017). The effects of attrition on the estimation of student outcomes: Evidence from UK universities. *Education Economics*, 25(3), 245-258. <https://doi.org/10.1080/09645292.2016.1258255>

Appendix

As illustrated earlier in the paper, multiple scenarios enables us to anticipate and prepare for potential future changes or disruptions, thereby mitigating risks and improving overall model performance.

1. Scenario 0:

In scenario zero, as demonstrated in the below figure and table, using the feature selection method with 22 variables 5 models have been used. The overall accuracy for the Random Trees 1 is % is 86.9 and the AUC is 0.88 with 26 fields resulted to be the best model in this scenario. While the other models were LSVM 1 accuracy is % is 0.84 and the AUC is 0.88 with 26 fields and the AUC with 0.84.

Field	Measurement	Values	Missing	Check	Role
STUDENT IDENTIFIER	Continuous	[7755837.0,8037098.0]		None	Input
STDNT_AGE	Ordinal	16.0,17.0,18.0,19.0,20...		None	Input
STDNT_GENDER	Flag	F,M		None	Input
STDNT_BACKGROUND	Nominal	"BGD 1","BGD 2","BGD...		None	Input
IN STATE FLAG	Nominal	"N,Y		None	Input
STDNT_MAJOR	Nominal	"Accounting,"Applied ...		None	Input
STDNT_MINOR	Nominal	"Accounting,"African ...		None	Input
FIRST_TERM	Ordinal	200508.0,200608.0,2...		None	Input
CORE COURSE_NAME 1 F	Nominal	"ANTH 1105","ANTH ...		None	Input
CORE COURSE_GRADE 1 F	Nominal	"A,B,C,D,F,INCOMPL..."		None	Input
CORE COURSE_NAME 2 F	Nominal	"ANTH 1105","ANTH ...		None	Input
CORE COURSE_GRADE 2 F	Nominal	"A,B,C,D,F,INCOMPL..."		None	Input
CORE COURSE_NAME 3 F	Nominal	"ANTH 1105","ANTH ...		None	Input
CORE COURSE_GRADE 3 F	Nominal	"A,B,C,D,F,INCOMPL..."		None	Input
SECOND_TERM	Ordinal	200602.0,200702.0,2...		None	Input
CORE COURSE_NAME 1 S	Nominal	"ANTH 1105","ANTH ...		None	Input
CORE COURSE_GRADE 1 S	Nominal	"A,B,C,D,F,INCOMPL..."		None	Input
CORE COURSE_GRADE 2 S	Nominal	"A,B,C,D,F,INCOMPL..."		None	Input
HOUSING_STS	Flag	"On Campus"/"		None	Input
RETURNED_2ND_YR	Flag	1,0/0,0		None	Target
DISTANCE_FROM_HOME	Ordinal	0.0,59.0,69.0,90.0,91...		None	Input
HIGH_SCHL_GPA	Continuous	[0.0,4.0]		None	Input
FATHER_HI_EDU_DESC	Nominal	College/Beyond,"High ...		None	Input
MOTHER_HI_EDU_DESC	Nominal	"College/Beyond,"Hig...		None	Input
DEGREE_GROUP_CD	Nominal	"A,B,V		None	Input
DEGREE_GROUP_DESC	Nominal	"Associate,Bachelors..."		None	Input
FIRST_TERM_ATTEMPT_HRS	Continuous	[9.0,21.0]		None	Input
FIRST_TERM_EARNED_HRS	Continuous	[0.0,21.0]		None	Input
SECOND_TERM_ATTEMPT_HRS	Continuous	[2.0,23.0]		None	Input
SECOND_TERM_EARNED_HRS	Continuous	[0.0,23.0]		None	Input
GROSS_FIN_NEED	Continuous	[0.0,2124900.0]		None	Input
COST_OF_ATTEND	Continuous	[0.0,2124900.0]		None	Input
EST_FAM_CONTRIBUTION	Continuous	[0.0,5999940.0]		None	Input
UNMET_NEED	Continuous	[-1212072.0,1632660...		None	Input
New Age	Ordinal	Equal 18,Less 18,Mor...		None	Input
Partition	Nominal	"1_Training","2_Testing"		None	Partition
SF-Factor-1	Continuous	[-4.64549653886138...		None	Input
SF-Factor-2	Continuous	[-3.70146470523567...		None	Input
SF-Factor-3	Continuous	[-2.56775008772755...		None	Input
SF-Factor-4	Continuous	[-1.64937623240398...		None	Input
SF-Factor-5	Continuous	[-3.03040712750744...		None	Input
SF-Factor-6	Continuous	[-13.9451730729230...		None	Input
SF-Factor-7	Continuous	[-5.4199501872063...		None	Input
SF-Factor-8	Continuous	[-4.01108620096217...		None	Input
SF-Factor-9	Continuous	[-2.89691849724089...		None	Input

Table 13. Variables Used for Scenario Zero

Use?	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift(Top 30%)	No. Fields Used	Overall Accuracy (%)	Area Under Curve	Accumulated Accuracy (%)	Accumulated AUC
<input checked="" type="checkbox"/>	Random Trees 1	< 1	1380.586	16	2.632	26	86.939	0.887	86.939	0.887
<input checked="" type="checkbox"/>	LSVM 1	< 1	1,060.0	8	2.371	26	84.988	0.84	84.988	0.84
<input checked="" type="checkbox"/>	Tree-AS 1	< 1	845.091	8	2.127	14	83.747	0.761	83.747	0.761
<input checked="" type="checkbox"/>	CHAID 1	< 1	843.571	8	2.070	15	83.747	0.765	83.747	0.765
<input checked="" type="checkbox"/>	XGBoost Tree 1	< 1	735.0	12	2.023	26	78.723	0.719	78.723	0.719

Figure 22. List of Models Used for Scenario Zero

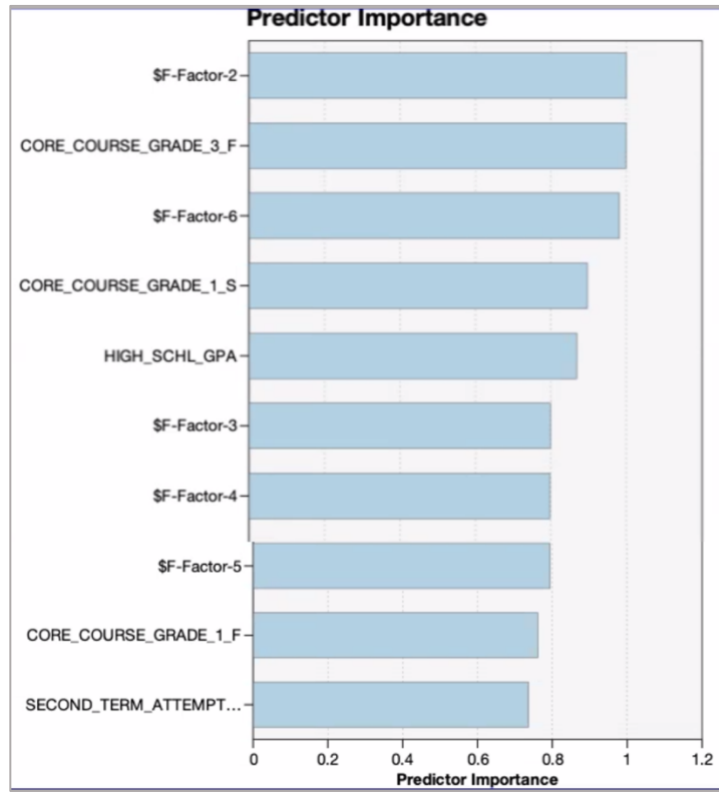


Figure 23. Predictor Importance for Random Tree 1 in Scenario Zero

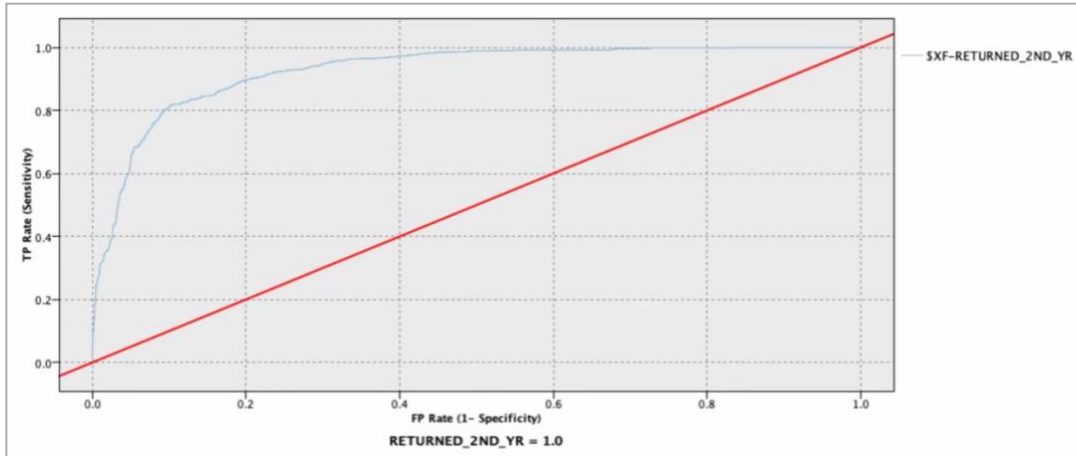


Figure 24. ROC for Random Tree 1 in Scenario Zero

According to Jayawant N. Mandrekar (2010), an illustration of a diagnosis test's sensitivity against its 1 -particularity is called ROC curve. The many cutpoints utilized for deciding whether the test outcomes are acceptable are represented by every point along the curve. The average test sensitivity across all potential specificity values, or simply vice versa, is represented to be as ROC curve.

2. Scenario 1:

In scenario one, as demonstrated in the below figure and table, using the feature selection method with 10 variables 5 models have been used. The overall accuracy for the LVSM 1 is 85.382% and the AUC is 0.723 with 10 fields resulted to be the best model in this scenario.

Field	Measurement	Values	Missing	Check	Role
☐ CORE COURSE GR...	♣ Nominal	"" ,A,B,C,D,F,INC...		None	☒ Input
☐ CORE COURSE GR...	♣ Nominal	"" ,A,B,C,D,F,INC...		None	☒ Input
☐ CORE COURSE GR...	♣ Nominal	"" ,A,B,C,D,F,INC...		None	☒ Input
☒ HIGH SCHL GPA	☒ Continuous	[0.0,4.0]		None	☒ Input
☒ FIRST TERM ATTE...	☒ Continuous	[9.0,21.0]		None	☒ Input
☒ FIRST TERM EARN...	☒ Continuous	[0.0,21.0]		None	☒ Input
☒ \$F-Factor-1	☒ Continuous	[-4.645496538...		None	☒ Input
☒ \$F-Factor-4	☒ Continuous	[-1.649376232...		None	☒ Input
☒ \$F-Factor-5	☒ Continuous	[-3.030407127...		None	☒ Input
☒ \$F-Factor-6	☒ Continuous	[-13.94917307...		None	☒ Input
☒ RETURNED 2ND YR	☒ Flag	1.0/0.0		None	☑ Target

Table 14. Variables used for Scenario One











Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in	Lift (Top 3...	No. Fields Used	Overall Accuracy	Area Under Curve	Accumulate Accuracy (%)	Accumulate AUC
<input checked="" type="checkbox"/>		 LSVM 1	< 1	215.0	6	1.958	10	85.382	0.723	85.382	0.723
<input checked="" type="checkbox"/>		 Logistic regression 1	< 1	210.0	5	1.957	10	84.823	0.729	84.823	0.729
<input checked="" type="checkbox"/>		 Tree-AS 1	< 1	205.373	5	1.95	4	84.358	0.714	84.358	0.714
<input checked="" type="checkbox"/>		 CHAID 1	< 1	215.0	6	1.939	4	85.102	0.716	85.102	0.716
<input checked="" type="checkbox"/>		 Neural Net 1	< 1	215.0	6	1.908	10	85.568	0.724	85.568	0.724

Figure 25. List of Models Used for Scenario One

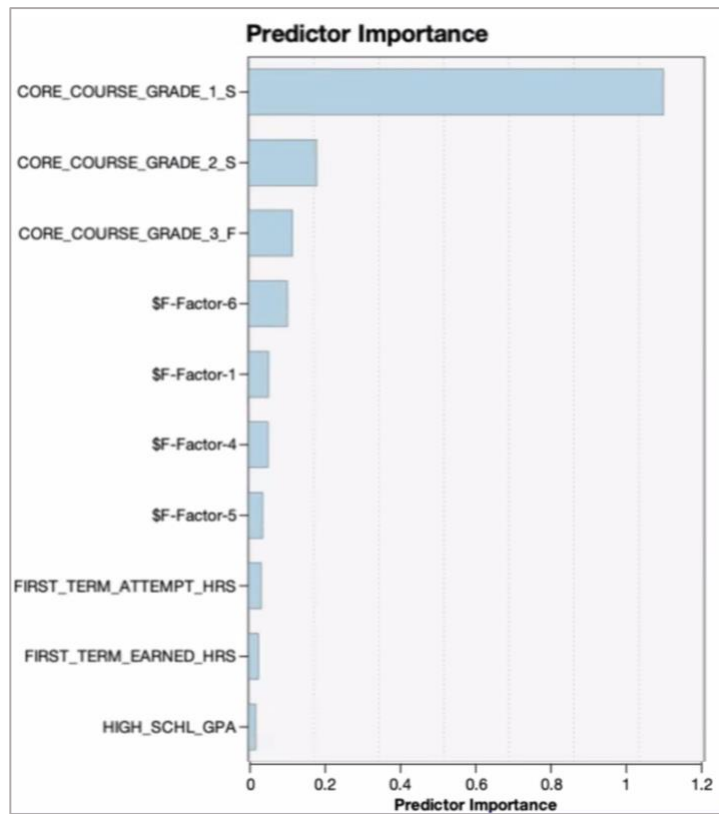


Figure 26. Predictor Importance for LVSM 1 in Scenario One

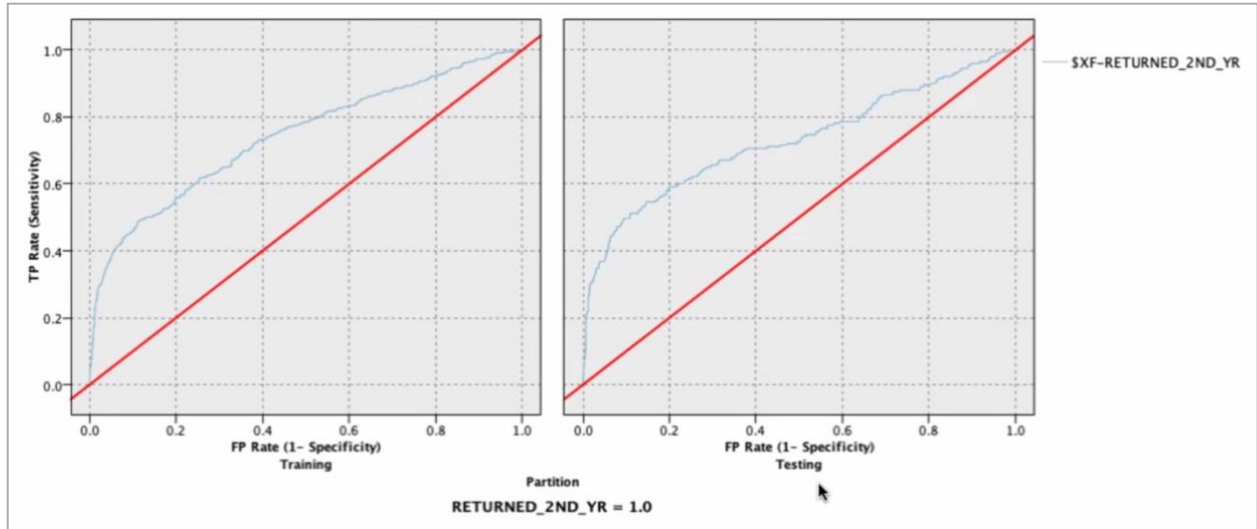


Figure 27. ROC for LVSM 1 in Scenario One

3. Scenario 2:

In scenario two, as demonstrated in the below figure and table, using the feature selection method with 10 variables 5 models have been used. The overall accuracy for the Random Tree 1 is % is 86.31 and the AUC is 0.885 with 21 fields resulted to be the best model in this scenario.

Field	Measurement	Values	Missing	Check	Role /
⊗ FIRST TERM	▮ Ordinal	200508.0,200...		None	↘ Input
⊠ CORE COURSE NAME 1 F	♣ Nominal	"" , "ANTH 1105" ...		None	↘ Input
⊠ CORE COURSE GRADE 1 F	♣ Nominal	"" , "A,B,C,D,F,INC...		None	↘ Input
⊠ CORE COURSE NAME 2 F	♣ Nominal	"" , "ANTH 1105" ...		None	↘ Input
⊠ CORE COURSE GRADE 2 F	♣ Nominal	"" , "A,B,C,D,F,INC...		None	↘ Input
⊠ CORE COURSE NAME 3 F	♣ Nominal	"" , "ANTH 1105" ...		None	↘ Input
⊠ CORE COURSE GRADE 3 F	♣ Nominal	"" , "A,B,C,D,F,INC...		None	↘ Input
⊗ SECOND TERM	▮ Ordinal	200602.0,200...		None	↘ Input
⊠ CORE COURSE NAME 1 S	♣ Nominal	"" , "ANTH 1105" ...		None	↘ Input
⊠ CORE COURSE GRADE 1 S	♣ Nominal	"" , "A,B,C,D,F,INC...		None	↘ Input
⊠ CORE COURSE GRADE 2 S	♣ Nominal	"" , "A,B,C,D,F,INC...		None	↘ Input
⊗ HIGH SCHL GPA	▮ Continuous	[0.0,4.0]		None	↘ Input
⊗ FIRST TERM EARNED HRS	▮ Continuous	[0.0,21.0]		None	↘ Input
⊗ SECOND TERM ATTEMPT HRS	▮ Continuous	[2.0,23.0]		None	↘ Input
⊗ SECOND TERM EARNED HRS	▮ Continuous	[0.0,23.0]		None	↘ Input
⊗ \$F-Factor-1	▮ Continuous	[-4.64549653...		None	↘ Input
⊗ \$F-Factor-2	▮ Continuous	[-3.70146470...		None	↘ Input
⊗ \$F-Factor-3	▮ Continuous	[-2.56775008...		None	↘ Input
⊗ \$F-Factor-4	▮ Continuous	[-1.64937623...		None	↘ Input
⊗ \$F-Factor-5	▮ Continuous	[-3.03040712...		None	↘ Input
⊗ \$F-Factor-6	▮ Continuous	[-13.9491730...		None	↘ Input
⊗ RETURNED 2ND YR	⚑ Flag	1.0/0.0		None	⊙ Target

Table 15. List of Variables used for Scenario Two

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in	Lift(Top ...	No. Fields Used	Overall Accuracy	Area Under	Accumulate Accuracy (%)	Accumulate AUC
<input checked="" type="checkbox"/>		Random Trees 1	< 1	1282.556	17	2.629	21	86.318	0.885	86.318	0.885
<input checked="" type="checkbox"/>		Logistic regression 1	< 1	935.0	12	2.25	21	84.043	0.821	84.043	0.821
<input checked="" type="checkbox"/>		LSVM 1	< 1	930.0	10	2.246	21	83.983	0.817	83.983	0.817
<input checked="" type="checkbox"/>		CHAID 1	< 1	784.000	8	1.991	8	83.392	0.739	83.392	0.739
<input checked="" type="checkbox"/>		XGBoost Tree 1	< 1	560.0	8	1.954	21	78.723	0.701	78.723	0.701

Figure 28. List of Models Used for Scenario Two

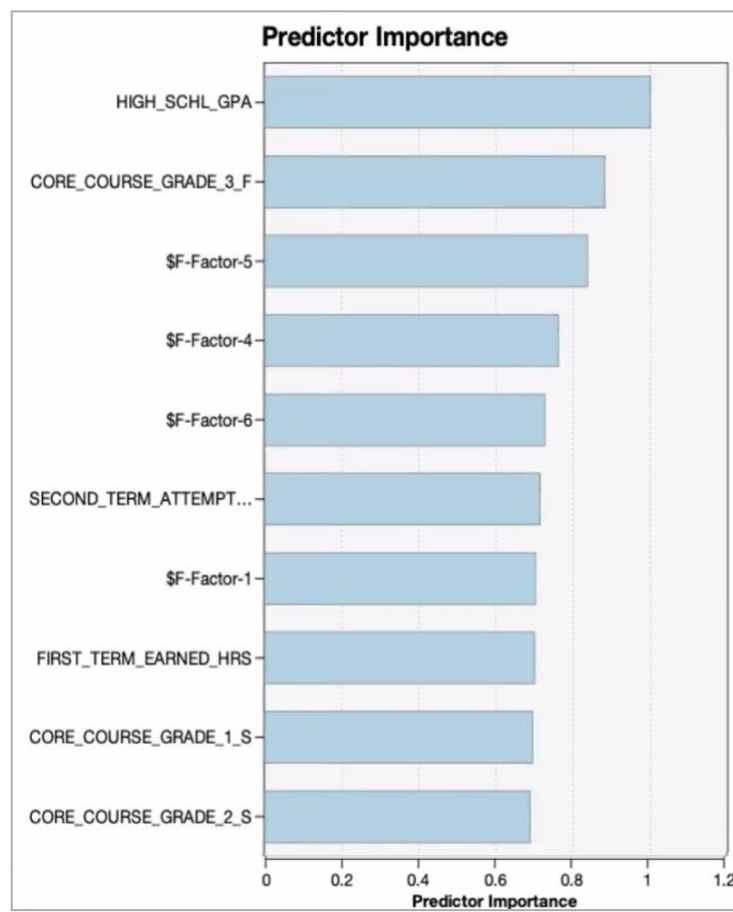


Figure 29. Predictor Importance for Random Tree 1 in Scenario Two

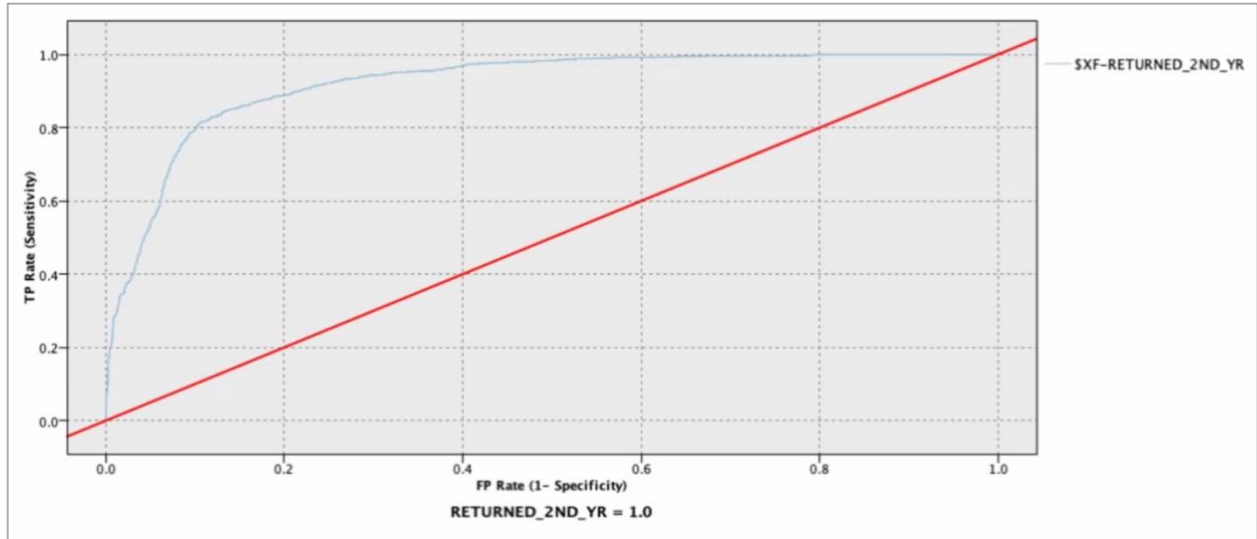


Figure 30. ROC for Random Tree in Scenario Two

Overall, we have demonstrated all the three different scenarios as shown in the below table. As a result, the best scenario among the three of them is scenario three. Also, SVM 1 with 90.928 and the AUC is 0.925 with 21 fields resulted to be the best model in this scenario.

Scenario No.	Best Model	Accuracy	AUC
Scenario 0	Random Trees 1	86.939	0.887
Scenario 1	LVSM 1	85.382	0.723
Scenario 2	Random Trees 1	86.318	0.885
Scenario 3	SVM 1	90.928	0.925

Table 16. Demonstrating All the Scenarios Used While Building the Models