AFRICAN CENTRE OF EXCELLENCE IN DATA SCIENCE
COLLEGE OF BUSINESS & ECONOMICS

STUDENT PERFORMANCE PREDICTION BASED ON MACHINE LEARNING ALGORITHMS AT THE ADVENTIST UNIVERSITY OF CENTRAL AFRICA

By

UZAYISENGA Emmanuel

Reg No: 221000424

A dissertation submitted in partial fulfillment of the requirements

For degree of Master of Data Science in Data Mining

African Centre of Excellence in Data Science,

College of Business and Economics,

University of Rwanda

Research Supervisor:

Assoc. Prof. Joseph NZABANITA

Kigali, July 2023

DECLARATION

I, UZAYISENGA Emmanuel Reg. No: 221000424, I hereby certify that this dissertation is the product of my efforts and has not been submitted for consideration for any other degree at the University of Rwanda or at any other institution. It was created and presented to the African Centre of Excellence in Data Science, College of Business and Economics. The dissertation has been reviewed by the administration, deemed to meet the established requirements by Turnitin's anti-plagiarism checker, and authorized.

Signature:

Printed Name: UZAYISENGA Emmanuel

July, 2023

APPROVAL

This dissertation entitled **STUDENT PERFORMANCE PREDICTION BASED ON MACHINE LEARNING ALGORITHMS at the Adventist University of Central Africa** written and submitted by UZAYISENGA Emmanuel in partial fulfillment of the requirements for the degree of Master in Data Science majoring in Data Mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 17 % which is less than 20% accepted by the African Centre of Excellence in Data Science (ACE-DS).

Signature:

Supervisor: Assoc. Prof. Joseph NZABANITA

DEDICATION

This thesis is dedicated to: My beloved wife BYUKUSENGE Honorine and my two children SHIMWA Emma Bianca and IGABE Gian Bevon for their immeasurable love and support.

This work is dedicated

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and acknowledge the following individuals and organizations who have contributed to the completion of my thesis:

I am immensely grateful to my supervisor Assoc. Prof. Joseph NZABANITA for his guidance, expertise, and unwavering support throughout the entire thesis process. Their valuable insights, constructive feedback, and encouragement were instrumental in shaping my research and improving the quality of my work.

I am grateful to my fellow graduate students, classmates, and colleagues who provided me with support, engaging discussions, and helpful suggestions throughout the course of my thesis work. Their collaboration and camaraderie created a conducive environment for intellectual growth and exploration. Family and Friends: I want to sincerely thank my family, friends, and loved ones for their continuous support, love, and understanding. My perseverance and inspiration were greatly aided by their unwavering encouragement, faith in my abilities, and support during trying times.

I would like to express my sincere gratitude to Mr. Charles NZARAMYIMANA for his unwavering support. Additionally, I extend my appreciation to the African Centre of Excellence in Data Science for providing the indispensable infrastructure, resources, and opportunities that have greatly facilitated my research and academic pursuits. Please know that your support and assistance are not only recognized but also deeply cherished. I am truly indebted to your invaluable contributions that have shaped my academic endeavors.

TABLE OF CONTENTS

DEDICATION ..... **III**

ACKNOWLEDGEMENTS ..... **IV**

LIST OF FIGURES ..... **VIII**

TERMINOLOGY ..... **X**

ABSTRACT ..... **XI**

CHAPTER 1: GENERAL INTRODUCTION **1**

1.1 BACKGROUND OF THE STUDY AND PROBLEM MOTIVATION 2

1.2 PROBLEM STATEMENT 3

1.3 RESEARCH MOTIVATION 4

1.4 RESEARCH CONTRIBUTION 4

1.5 OBJECTIVE OF THE STUDY 5

1.5.1 General objective *5*

1.5.1 Specific Objectives *5*

1.6 RESEARCH QUESTIONS 5

1.7 THE SCIENTIFIC KNOWLEDGE TO BE GAINED FROM THE THESIS 6

1.8 RESEARCH HYPOTHESES 6

1.9 SIGNIFICANCE OF THE STUDY 6

1.10 SCOPE OF THE WORK 6

1.11 OUTLINE 7

CHAPTER 2: LITERATURE REVIEW **8**

2.1 INTRODUCTION TO PREDICTING STUDENT PERFORMANCE 8

2.2 CONCEPTUAL REVIEW 8
2.2.1 Academic GPA 8
2.2.2 Socio-economic 8
2.2.3 Principle passes can Influence Students' Academic Performance 9
2.2.4 Sponsorship can Influence Students' Academic Performance 9
2.3 THEORETICAL REVIEW 10
2.4 EMPIRICAL REVIEW 11
2.2.5 Related Studies on Predicting Student Performance 11
2.5 GAPS IN THE LITERATURE 12
CHAPTER 3: METHODOLOGY 13
3.1 Research Design 13
3.2 Population of the Study 13
3.3 Data Preprocessing 13
3.4 Data Preparation 14
3.5 Classification Techniques for Predicting Student Performance 15
3.6 Factors Influencing Student Performance 16
3.7 Research Method Description 16
3.8 Study Approach 17
3.10 Data Mining 19
3.11 Time Series Data 21
3.12 Data Leakage 21
3.13 Data Imbalance 22
3.14 APPLICABLE ALGORITHMS 23

CHAPTER 4: RESULTS..... 26

4.1 DATASET DESCRIPTION 26

4.2 EDA (EXPLORATORY DATA ANALYSIS)..... 27

4.3 STUDENTS IN PROHIBITION 31

4.4 STUDENTS WITH TWO PROBATIONS. 35

4.5 LINEAR REGRESSION 37

4.6 DECISION TREE REGRESSION..... 38

4.7 RANDOM FOREST REGRESSION..... 40

4.8 LASSO..... 43

4.9 DISCUSSION 46

CHAPTER 5: CONCLUSION..... 49

REFERENCES..... 50

LIST OF FIGURES

Figure

1. The element of the confusion matrix 18
2. Figure is an illustration of ML classification 21
3. The graph which is showing GPA of all registered students 2019-2020 semester 1 and 2 26
4. GPA Eamed by Students Regarding their Departments 27
5. Students' Evaluations on Their Faculty's GPA 28
6. shows the students in the different countries 29
7. GPA Earned by Students with respect their Gender 30
8. Students in probation by country 31
9. Students in probation by faculty 32
10. Financial status of students in probation 33
11. Principle pass of students in probation 34
12. Students in probation by gender 35
13. Students in probation by facilty and country 36
14. Actual Predicted by Lasso ..... **Error! Bookmark not defined.**
15. Shows the factors influencing student performance 45

LIST OF TABLES

Table

1. Showing the dataset attribute 15

2. Difference models tests results. 44

TERMINOLOGY

AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
AUC	Area under the curve
CV	Computer Vision
GPA	Grade Point Average
GPU	Graphics Processing Unit
KNN	K-Nearest Neighbor
ML	Machine Learning
MLT	Machine Learning Techniques
NLP	Natural language processing
RF	Random forest
SVM	Support Vector Machine

ABSTRACT

In recent years, there has been a growing interest in predicting student performance in educational institutions using machine learning. This paper presents an approach to predict student performance based on data mining classification techniques. The objective was to develop a predictive model that can accurately forecast student performance and identify factors that influence academic success. The proposed approach involves the collection of various attributes related to students, such as identification numbers, department, faculty, GPA earned, nationality, place of birth, religion, gender, student absence days, parents' school satisfaction, and principal pass. These attributes are used as input variables for classification algorithms, including decision trees, logistic regression, and support vector machines. The dataset used in this study is obtained from University in the office of registrar and consists of historical data of students, including their grades and other relevant information. The dataset is preprocessed to handle missing values, outliers, and categorical variables. The dataset has 25 features and 2583 observations which are enough for training our models in doing experiment. Different classification algorithms are applied to the preprocessed dataset, and their performance is evaluated using various evaluation metrics, such as accuracy, precision. The experimental results indicate that the proposed approach achieves promising performance in predicting student performance. The decision tree algorithm outperforms other classifiers with the highest accuracy. The selected features reveal that factors such as previous academic performance, identification numbers, department, faculty, GPA earned. Influence student performance. Findings showed that the decision tree predicts student performance with accuracy of 99.87%. By leveraging student-related attributes and employing machine learning algorithms, educational institutions can enhance their decision-making processes and provide targeted support to students, ultimately improving overall educational outcomes.

Keywords: Academic performance, Machine Learning

CHAPTER 1: GENERAL INTRODUCTION

Predicting student performance has become a critical aspect in the academic environment, and it plays an important role in producing high-quality graduates (Joazeiro, 2021)

Universities generate large volumes of data with reference to their students in traditional paper and electronic forms (Basar1, Mansor, & Jamaludin, 2021). In recent time, the advances in the data mining field make it possible to mine these Universities educational data and generate information that provide and assist teachers and students in decision making.

Not just administrators, academicians, and educators should be deeply concerned about students' success at universities, but also the government and their parents (Bhardwaj & Pal, 2021). Academic achievement is one of the main criteria the looks at when hiring flesh graduates. As a result, to fulfill the requirements of the employer, students must exert the most effort in their academic work.

Every higher education facility must also have a database where all pertinent information about the academic activity is stored and maintained. A repository of student results is a type of educational database. It is a big data bank that holds students' ungraded grades for various courses. (Junshuai, 2019)

The university administration wants to determine which data attributes are the best predictors of university achievement. They'd also like to know if the data collected is sufficient for producing valid predictions, if any changes to the data collection method are necessary and how to improve it, and what more data to collect to make the analytic results more usable.

In their first year of computer science student’s performance can make or break their educational trajectory, and it typically has a significant impact on their GPA.

The evaluation criteria for the students, such as midterm and final exams, assignments, and lab work, are examined. Before the final exam is given, it is advised that the class teacher be informed of all this related material.

The results of this study will assist teachers in raising student achievement and significantly lowering the dropout rate. In this research, we provide a hybrid method based on Linear Regression, Decision Tree Regression Random Forest Regression, SVR and Lasso Regression that enables academics to forecast student GPA, and based on that, instructors can take the necessary steps to enhance student academic performance.

A popular measure of academic achievement is the grade point average (GPA). Many academics universities have established a minimum GPA that should be kept. Thus, GPA continues to be the most important element utilized by academic planners to identify weak students and aid administration in initiating corrective action to generate great graduates who would graduate at least with the upper second category.

1.1 Background of the Study and Problem Motivation

In the first year of their university studies, several students leave their programs. (Willging J. , 2018) Found that 29% of Sweden's full-time bachelor's students enrolled in the courses dropped out. There are several potential causes for this phenomenon, with motivation and progress perception being two of the main ones proposed in this study as being causal. The project also looks into ways to give helpful criticism and, in a sense, turn the learning process into a game to help students finish their university degrees more frequently. (Kowsari, MeimandI, Heidarysafa, & Mendu, 2019)

Machine learning (ML) can be used to investigate this pattern of behavior. This is undoubtedly a wide topic, and there are numerous issues and difficulties to be resolved. The broad category of categorization is the focus of this investigation. Classification in machine learning (ML) is the process of figuring out which category an observation belongs to, according to (Soofi A. A., 2017). In essence, it's a process that develops correlations between an independent variable (which might be categorical or numerical in type) and a dependent variable (which is categorical in nature).

There are numerous decision-making issues that fall under the classification area right now. According to one study, there are two empirical learning strategies for classifying this kind of problem. Statistical Pattern Recognition (SPR) and Machine Learning Techniques (MLT), the

latter of which builds decision trees and production rules, are the first and second, respectively. The latter type will be examined in this study together with other MLTs. (Romero & Ventura, 2013)

1.2 Problem Statement

It is well known that students are leaving various bachelor's degree programs at universities at a particular pace.

The same case in AUCA, where students who failed to get 12/20 basic scores twice are forced to change the faculty and department to study. For example, student in computer science who failed twice should change to business admiration or Education. If he failed again, he/she will have fired in the school or drop out him/herself.

There are several causes for this, but for the sake of this study, motivation and perception of progress are the key areas of interest. Technically, however, it's more interesting to compare many ML models employed in ML frameworks in order to reach agreement on the model that is more accurate in predicting student success.

This study will assess five machine learning models within machine learning frameworks. It will utilize student data from secondary education and the first-year university grades to conduct a comparative analysis of various ML models, including Linear Regression, Decision Tree Regression, Random Forest Regression, SVR, and Lasso Regression. The global impact of this issue on students and universities underscores its significance. The primary aim of this thesis is to explore the potential of using student data and machine learning to identify individuals in need of academic support during their university studies.

1.3 Research motivation

Overall, the motivation behind Student performance prediction based on machine learning algorithms lies in improving educational outcomes, providing personalized support, optimizing resource allocation, and fostering evidence-based decision-making. By leveraging the power of data analysis and machine learning, educational institutions can enhance student success, promote equity, and continuously improve the quality of education.

Predicting student performance can aid in optimizing resource allocation by identifying which students require additional support and allocating resources accordingly. This ensures that resources are efficiently utilized to enhance the educational experience and outcomes for both struggling students and high achievers. Every student is unique, with individual strengths, weaknesses, and learning styles. Data mining classification techniques can help identify these individual characteristics and tailor educational interventions accordingly. By analyzing student data, such as past academic records, demographics, and learning behaviors, personalized recommendations can be generated to enhance the learning experience and maximize student performance.

Evidence-Based Education: Student performance prediction based on machine learning algorithms promotes evidence-based decision-making in education. By leveraging data-driven insights, educational institutions can move away from subjective judgments and make informed decisions backed by empirical evidence. This promotes accountability, transparency, and efficiency in the education system.

1.4 Research Contribution

This study adds to the corpus of knowledge by predicting student academic performance and assisting in the identification of students with bad grades who can then be examined and given new materials and strategies to help them improve their marks. Predicting a student's performance allows an instructor to identify non-engaged students based on their class attendance, class materials and online learning platform actions and activities. It also aids in identifying problematic students and provides teachers with a proactive opportunity to devise additional methods to boost their chances of passing during their study program.

1.5 Objective of the Study

1.5.1 General objective

The objective of this research is to train a machine learning model, selecting from decision trees, random forests, support vector machines, neural networks, and Lasso, with the goal of achieving a high level of accuracy in predicting student performance.

1.5.1 Specific Objectives

The specific objectives of the study are:

1. To implement a predictive model that estimates students' future grades based on their historical performance and other relevant factors.
2. To train a machine learning algorithm on a dataset of 2853 student records with associated performance indicators.
3. To train a machine learning model that accurately classifies students into predefined performance levels.

1.6 Research Questions

This main research question is dissected into the subsequent questions:

I. What is the optimal way to apply ML models (Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression [SVR]), and which model exhibits superior accuracy and precision when applied to the university's student dataset?

II. To what extent is a dataset required to draw reliable conclusions for the classification of the dataset?

III. Can ML algorithms serve as a suitable method for identifying patterns among students, and are they effective in this regard?

1.7 The scientific knowledge to be gained from the thesis

1. Assessing the suitability or validity of employing ML algorithms to forecast student performance solely based on datasets obtained from the registrar's office.

2. Examining the performance of Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR) under specific conditions to identify the model that excels in terms of accuracy, precision, and recall. Additionally, offering insights to enhance the performance of ML models when applied to this particular dataset.

1.8 Research Hypotheses

i. Null Hypothesis

There is no significant influence of student absence, student background, student sponsored and student principal passes on student performance when using machine learning models.

ii. Alternative Hypothesis

There is a significant influence of student absence, student background, student sponsored and student principal passes on student performance when using machine learning models.

1.9 Significance of the Study

Overall, this study on predicting student performance has significant implications for improving student success, promoting educational equity, optimizing resource allocation, and fostering evidence-based decision-making in education. It empowers educators, administrators, policymakers, and researchers with valuable insights to enhance educational practices, support student growth, and contribute to a more equitable and effective education system.

1.10 Scope of the work

The scope of this work typically was involving the following aspects;

Data Collection. The first step was to gather relevant data related to student performance. This may include academic records, GPA earned, nationality, place of birth, religion, gender, student

absence days, parents' school satisfaction, previous exam scores, socioeconomic factors, and any other relevant variables.

The research also was focused on Data Preprocessing, Feature Selection/Extraction, Classification Techniques, Model Development and Evaluation, Interpretation and Insights

1.11 Outline

In order to better follow along with the content that will be presented further in the thesis, it is helpful to understand the theory and a brief description of a few technologies, algorithms, and ideas related to machine learning (ML) in Chapter two. The general concept of preventive maintenance is introduced in Chapter 1's introduction. Along with the advantages of such a system to its end users, the primary goal of the research and its aims are described. The problem for which we are looking for a solution and the scope are both stated explicitly throughout the chapter.

In chapter three, it is explained how the scientific method have applied to this study, how the thesis will be brock down into different steps to be carried out, and how the thesis evaluation process will be carrying out. Chapter four discusses the various options that will be taken into consideration (In this instance, the framework choice) for the experiment in the study and explains why python language was selected as the framework to be used.

The experiment's results will be presented in Chapter four as tables, plot diagrams, and bar charts. Chapter five will offer a summary of the study, the conclusions reached and the reasoning behind them, as well as some recommendations for further research, and it will conclude with some ethical and societal problems to be considered.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction to Predicting Student Performance

In recent years, there has been growing interest in utilizing data mining classification techniques to predict student performance in educational settings. Predicting student performance has significant implications for educational institutions, as it enables timely interventions and personalized support to improve student outcomes. However, the complex and multidimensional nature of student performance necessitates the use of advanced analytical approaches. This literature review aims to explore the existing research on Student performance prediction based on machine learning algorithms and identify gaps for further investigation.

2.2 Conceptual Review

I must first identify the potential variables that might have an impact on the students grades and behavior in order to forecast their academic achievement. Academic GPA in the background study, (Njuguna, 2021)such us Sponsorship, Principle passes in secondary schools, social problems can also be influencing the performance of the students

2.2.1 Academic GPA

According to ISLJAMOVIC and SUKNOVIC (2014), the grade point average (GPA) is an important factor in determining a student's overall past academic accomplishments and future potential for a variety of purposes, including college admission, graduate program admission, scholarship awards, and entry into training programs and the workforce. The GPA is sometimes considered first for these objectives even if a range of other measurements and outcomes may also be taken into consideration. This is because it is believed to reflect a student's aptitude and potential for the future in a straightforward, numerical, and simply comparable manner. (Volwerk & Tinda, 2021).

2.2.2 Socio-economic

Economic and social factors have been identified as key factors influencing students' performance in national exams. This research aimed to identify the socio-economic factors

impacting academic achievement in public primary schools within Murang'a South Sub County. The study utilized a descriptive survey design, incorporating both quantitative and qualitative methods. The research encompassed all public primary schools in Murang'a South Sub County, Kenya, involving deputy head teachers, teachers, and pupils as part of the study population (Njuguna, 2021).

2.2.3 Principle passes can influence students' academic performance

According to Rulinda, Ephrard; Role, Elizabeth (2013), The best legacy a country can leave its children is education. This would imply that a country's ability to progress depends heavily on the caliber of its educational system. Most people agree that developing human resources must come first in order for any real progress to take root. In any community, formal education continues to be the principal means of promoting social mobilization and socioeconomic growth (Bussolo, Checchi, & Peragine, 2019).

The General Academic Regulations outline the general academic requirements for admission. For admission to the first year of an undergraduate program, a minimum requirement includes an Advanced General Certificate of Secondary Education with at least two principal passes, enabling entry into higher education. Alternatively, candidates can provide a qualification or other evidence demonstrating their capability to engage in the program, deemed equivalent. Additionally, a demonstration of English language proficiency at the level required for level 1 higher education is mandatory. It's important to note that specific programs may have requirements exceeding the stated minimum criteria (Rulinda & Role, 2013).

2.2.4 Sponsorship can influence students' academic performance

Studies have indicated that children who have their families actively involved in their education Perform well on tests, attend school more frequently, turn in more homework, and exhibit more positive attitudes and behaviors. (Henderson and Berla, 2021)

In this paper, a management viewpoint was used to determine the impact of sponsorship on secondary school students' academic achievement. The researcher evaluated the academic performance of pupils in Compassion International-funded initiatives in the Ndeiya division using the KCSE test as the benchmark. The donors provided the students with spiritual support

in the form of qualified teachers providing spiritual nourishment on Saturdays and Sundays, opportunities to participate in youth-focused church programs, encouragement to attend church services, guidance and counseling, instruction in morality, and the distribution of spiritual and personal development books like the Bible and hymnals. (NDUNGI, 2012).

The kind of social support given by donors to sponsored children includes meeting their basic needs, such as providing food and clothing to them and their families, monitoring them while they are in school to ensure successful learning, and taking them on trips and excursions to inspire them and improve their interaction. The sponsored children received physical support in the form of adequate housing for the family, water tanks and other amenities needed to raise living standards, game kits, and medical/health assistance in the form of payment of hospital bills for the kids. The academic success of the sponsored youngsters was enhanced by their compassionate sponsorship. In order to regulate the effects of sponsorship on secondary school academic performance in Kenya, the study suggests reviewing the current policies and legislation on the compassion sponsorship program. This will fill any gaps in the regulations and policies governing the execution of the sponsorship program and improve their efficacy. (NDUNGI, 2012)

2.3 Theoretical review

Machine learning (ML) can be a difficult topic to address because it encompasses a sizable subset of research and studies on numerous algorithms with the goal of tackling challenging tasks and patterns. This chapter will discuss a number of concepts and algorithms that the reader should be familiar with before moving on to the next chapter. (Chavez, 2021)

In this study, we set out to develop an accurate prediction model based on only behaviors that could be affected by intervention, and only the earliest of these behaviors, to ensure there was time in the semester to provide learning support and observe effects of this treatment. The behavior-based prediction model successfully identified 75% of those who would need to repeat the course; when that model was applied and a digital advice intervention was deployed to learners with similar behaviors (Bernacki & Uesbeck, 2020)

2.4 Empirical review

Empirical research is characterized as any investigation in which the study's conclusions are derived exclusively from concrete empirical evidence, ensuring its verifiability. This empirical evidence can be collected through both quantitative market research and qualitative market research methods.

2.2.5 Related Studies on Predicting Student Performance

A comprehensive review of the literature reveals several studies that have applied data mining classification techniques to predict student performance. For instance, (Mesaric, 2016) used decision trees to predict students' final grades based on variables such as previous exam scores, attendance, and study habits. Their findings demonstrated the potential of classification techniques in accurately identifying at-risk students. Similarly, (Brodie, June2029)employed logistic regression to predict dropout risk among college students using demographic data, academic records, and social economic factors. These studies highlight the diverse applications and promising results of data mining classification techniques in predicting student performance (Korde, 2012).

Another study, titled "Student Performance Evaluation in Educational Data Mining," involved the development and assessment of two machine learning models by the researchers (Pena, 2016). The initial model was an Artificial Neural Network (ANN), and the second was a Random Forest (RF), both utilizing TensorFlow as their backend. The objective was to predict students' academic performance by analyzing evaluation data and geographic information. The researchers aimed to determine if ANN, known for its performance in various data domains, could excel when applied to a substantial amount of educational data. Furthermore, the study aimed to compare the effectiveness of ANN with that of RF.

The increasing interest among academics in employing data mining to analyze student data has inspired the undertaking of this research. According to the study's authors, these advancements in data mining hold potential benefits for the educational sector. The authors concluded in the final chapter that recurrent neural networks (RNN) could play a future role in identifying

students on the verge of dropping out or withdrawing from classes. This study aims to build upon the existing research by addressing this research gap or exploring future work in this area.

2.5 Gaps in the Literature

While significant progress has been made in Student performance prediction based on machine learning algorithms, there are still notable gaps in the literature. Some areas that require further investigation include:

1. Limited studies on the application of advanced classification techniques, such as deep learning and ensemble methods, in predicting student performance.
2. Insufficient exploration of the impact of contextual factors, such as school characteristics and teaching methodologies, on student performance prediction.
3. Inadequate focus on longitudinal analysis and prediction, which can provide insights into student trajectories and long-term performance outcomes.
4. Limited research on the interpretability and explain ability of predictive models, which is essential for stakeholders in the educational domain to understand and trust the predictions.

CHAPTER 3: METHODOLOGY

3.1 Research Design

According to Leedy and Ormrod (2015), a research design is a carefully set of plans developed by a researcher that lays out requirements and guidelines for the study or research. According to Gay et al. (2006), a research establishes and reports the state of things; it entails gathering numerical data to test hypotheses or respond to inquiries regarding the present condition of the study's subject. The present study's design was appropriate because it helped explain how several elements, including sponsorship, background GPA, teachers, student absence, and instructional leadership, influence academic success.

3.2 Population of the study

This study makes use of first-year student behavior data from all AUCA departments. The information for this study was gathered through qualitative class observations that took place over the course of two semesters. The population was consisting 2583 students who were enrolled in academic year 2020-2021 Semester 1 and 2020-2021 Semester two. The dataset of this study involves the collection of various attributes related to students, such as identification numbers, department, faculty, GPA earned, nationality, place of birth, religion, gender, student absence days, parents' school satisfaction, and principal pass.

3.3 Data Preprocessing

The present features are helpful but some of them were identified as personal data which can cause privacy breaches, yet they do not participate in our case study. Therefore, they were removed from dataset. Those features were the telephone number and the name of the student.

Secondary, in preparing the dataset some features were identified to provide no information about the prediction of GPA of a student. Those features have the unique data assigned to

every person which does not bring insights on the dataset. There they were dropped. Those features are the cardinal number assigned to every student and the StudentID.

In preparing the dataset, there are some values which were missing. Each feature was handled separately from each other because of different characteristics. Those columns with missing numbers were the faculty, GPA, RELCODE, DOB, Gender, RELIGION, Program, Sponsored or not sponsored, Announcements View, Principle Pass, Observation. Among these columns, Observation and Religion was identified as have high number of missing values above 50% which lead them to be dropped because if a variable has more than 50% missing values, you do not have enough information to ascertain anything useful.

In Handling missing values, each feature was identified as continuous or categorical. These helped in deploying appropriate technique in dealing with them. The repetition of the data in each feature was calculated and found that features with more than 10% of unique data compared to observations were taken as continuous data whereas those features with the unique data which are less than 10% compared to observations were taken as categorical. All features were identified as categorical features except there two features which are ‘GPA’ and ‘DOB’. In handling categorical features, the mode techniques, which is to identify the most frequent element in the features and then used to fill the missing values, was used. In handling continuous data features, firstly we must be aware of the distribution of the data in a feature.

3.4 Data Preparation

Data preparation is an imperative step of the Education data mining process. The collected data has been processed to remove any erroneous or missing data and transformed into a format that can facilitate maximum extraction of knowledge. **Table 1** presents attributes considered in this study.

Attributes	Description
StudID	Student Identification Number
StudName	Student names
Department	Department of the students
Faculty	Faculty of the students

GPAAll2019-2022 Sem1	Grade Point Average for all students in semester one in academic year 2019-2020
GPAAll2019-2022 Sem2	Grade Point Average for all students in semester two in academic year 2019-2020
Drop out2019 sem2	Drop out student enrolled in 2019 all semesters one and two
Nationality	Nationality of students
Placeofbirth	Place of birth of the student
RELCODE	Religion CODE of the students
DOB	Date of birth of the student
Gender	Gender
Religion	Religion
Program	Program wh
Sponsored or not sponsored	Students sponsored by sponsored or not
Tel-NUMBER	Telephone number of the student
StudentAbsenceDays	Attendance of the student
AnnouncementsView	Announcements View
ParentAnsweringSurvey	Parents Answering Survey
ParentschoolSatisfaction	Parents school Satisfaction
PrinciplePass	Principal passes from secondary school
Obsarvation	Observation

Table 1: Showing the dataset attribute

3.5 Classification Techniques for Predicting Student Performance

Several classification techniques have been utilized to predict student performance with varying degrees of success. (Feng, 2019) Decision trees offer interpretability and intuitive decision rules, while random forests combine multiple decision trees to improve prediction accuracy (Sharma, 2023). Support vector machines handle complex datasets by finding optimal hyper planes to separate classes. Logistic regression provides insights into the probability of student success or failure. Naive Bayes utilizes probabilistic calculations to predict class membership, and artificial neural networks capture complex relationships through interconnected layers of nodes. (Sharma, 2023). Comparative studies between these techniques have been conducted to identify the most effective approach for predicting student performance in different educational contexts.

3.6 Factors Influencing Student Performance

Mapuranga, Zebron and Musingafi (2015) Student performance is influenced by a multitude of factors, including academic GPA in the background study, socio-economic such us Sponsorship, Principle passes in secondary schools etc… Understanding these factors is crucial for developing accurate predictive models

And designing effective interventions to improve student outcomes. By incorporating relevant factors into the predictive models, researchers can gain a deeper understanding of the complex relationships that impact student performance.

3.7 Research method description

This study encompassed a total of five phases. The initial phase involved conducting a comprehensive literature review, examining research articles, surveys, journals, and e-books through Google Scholar. The purpose was to acquaint the reader with the current state-of-the-art and establish a research gap, providing justification for the ongoing research.

The second phase entailed the design and planning of the experiment. The student dataset obtained from the Registrar's office at Adventist University of Central Africa underwent analysis, considering various cleaning methods to prepare for machine learning (ML) training. Subsequently, different frameworks were evaluated for training and comparing ML algorithms. The study also explored diverse "scoring" methods to assess the trained ML models.

The third phase, the Implementation phase, closely aligned with the second phase. It involved executing the plan, encompassing tasks such as cleaning the student dataset using Excel, learning, and training the ML models in the Python framework, scoring the models with various metric measuring functions, and ultimately evaluating the models by generating metric values and graphs.

The fourth phase revolved around scoring, plotting, and comparing the data generated by the ML models. Scoring, synonymous with prediction, involved creating values from new input data using trained ML models, which could predict future values or represent a predicted category or event.

The fifth and final phase centered on analyzing the results derived from the measured data. Leveraging the Python framework, various tools and approaches were considered for data analysis. Selecting specific tools, the study visualized the models through diagrams and charts to draw meaningful insights from the results.

3.8 Study approach

This research used a quantitative methodology to address and respond to the research queries. The primary research question sought to investigate the appropriateness and efficacy of ML algorithms in discerning patterns within student data.

The method used to achieve this objective was general because the research question was similarly general. This objective was achieved by carrying out the study's main experiment, which involved cleaning the student data before feeding it into various ML models to produce a pattern, whether it was supervised (classification vs. regression) or unsupervised (clustering vs. dimension reduction vs. association). The study question was answered when a pattern appeared.

The second research question in this study aimed to implement ML models (SVM and ANN) and assess which model demonstrated superior performance in terms of accuracy and precision in Python, utilizing datasets from the university. The approach to address this question involved precisely defining parameters, measuring their values, and making comparisons. In this context, Accuracy was specifically defined as the proportion of predictions correctly made by an ML model (Machine Learning, 2022).

The third research question sought to determine the amount of information required (Brownlee, 2019) to classify datasets effectively and identify individuals in need of assistance versus those who did not. The research followed a specific approach in addressing this question.

In order to determine whether the data was enough, attained accuracy was compared to baseline accuracy. The amount of data utilized was regarded adequate if the accuracy gained by training the model was higher than the accuracy of the baseline data. By dividing the number of the

larger set by the total number of TRUE data points (TRUE = giving a value of "1" = cleared course), the baseline accuracy was determined.

3.9 Evaluation Metrics

Various metrics are used to assess how well machine learning models complete the tasks for which they were designed. Precision, recall, accuracy and the Area under the Curve (AUC) are a few of these measurements. The effectiveness of constructed models on unobserved data was assessed using the assessment metrics recall, precision score, ROC-AUC, and confusion matrix.

The confusion matrix serves as a frequently utilized performance assessment metric in machine learning classification assignments. It furnishes an intricate depiction of the classification model's performance, encapsulating the model's predictions and juxtaposing them with the actual class labels in the dataset.

A confusion matrix is typically a square matrix that has rows and columns representing the true and predicted classes, respectively. It helps in understanding the distribution of predictions across different classes and identifying potential errors or misclassifications made by the model. The matrix is constructed as follows:

The elements of a confusion matrix are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

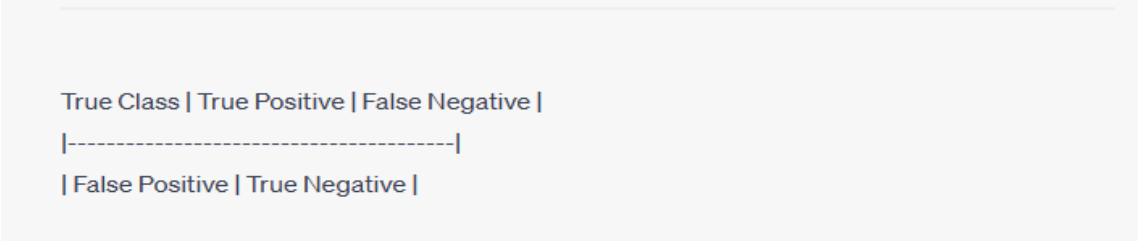


Figure 1 The element of the confusion matrix

a) True Positive (TP)

The model correctly predicted a positive class instance.

b) False Negative (FN)

The model incorrectly predicted a negative class instance when it should have been positive.

c) False Positive (FP)

The model incorrectly predicted a positive class instance when it should have been negative.

d) True Negative (TN)

The model correctly predicted a negative class instance.

Using the values from the confusion matrix, various performance metrics can be derived, including:

Accuracy

It represents the proportion of accurate predictions made by the classification model and can be computed using the following mathematical formula:

$$Accuracy = \frac{Number\ of correct\ predictions}{Total\ number\ of\ predictions}$$

$$OR$$

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

$$Where:$$

$$TP = True\ Positive, TN = True\ Negative,$$

$$FP = False\ Positive, FN = False\ Negative$$

Were as Precision were defined as the proportion of positive identifications and the proportion of actual positives respectively (Machine Learning crush caurses , 2021).

$$Precision = \frac{TP}{TP + FP}, \qquad Recall = \frac{TP}{TP + FN}$$

3.10 Data mining

Stedman (2021) says that data mining is a procedure for knowledge discovery that uses the right algorithms to analyze historical data to provide knowledge or insight that would otherwise be

concealed. Large datasets are mined for patterns through the data mining process, which helps people make decisions. Different types of problems, such as classification difficulties, regression problems, clustering problems, and prediction challenges, could be solved using data mining approaches. Either supervised learning or unsupervised learning could be used to complete a task.

3.9.1 Supervised learning

Supervised learning is a machine learning approach in which an algorithm learns from labeled training data to make predictions or infer patterns in unseen data. It involves training a model using input data that is already paired with corresponding target labels or desired outputs.
In supervised learning, the algorithm learns the relationship between the input features (also known as predictors or independent variables) and the corresponding target variable (also known as the dependent variable or label). The goal is to develop a model that can accurately generalize and make predictions on new, unseen data. (Liu & Wu, 2022)

3.9.2 Unsupervised learning

Siadati (2018) says that unsupervised learning is a machine learning approach in which an algorithm learns patterns, structures, or relationships in unlabeled data without explicit target labels or desired outputs. Unlike supervised learning, unsupervised learning does not require pre-labeled data for training.

In unsupervised learning, the algorithm explores the data to identify inherent patterns, groupings, or representations that may be present. The goal is to discover hidden structures or relationships within the data without any prior knowledge or guidance.

Unsupervised learning has various applications, including customer segmentation, market basket analysis, anomaly detection, image or text clustering, recommender systems, and data visualization. (Nur Yilmaz & Çimtay, 2018)
It's important to note that unsupervised learning can also be used in conjunction with supervised learning. Unsupervised learning techniques, such as pre-training or feature engineering, can help improve the performance of supervised learning models by extracting meaningful representations or reducing the complexity of the data.

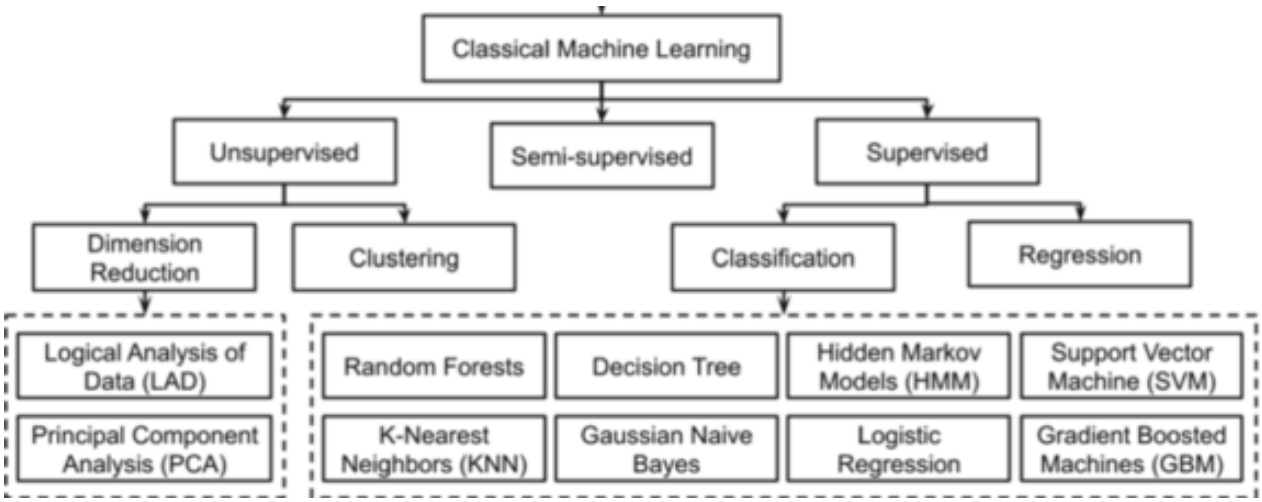


Figure 2: Figure is an illustration of ML classification

3.11 Time Series Data

Time series data refers to a collection of observations or measurements recorded at specific time intervals and arranged in chronological order. In time series data, the time component plays a critical role as it captures the sequential nature of the observations. This type of data is commonly encountered in various fields, including finance, economics, weather forecasting, stock market analysis, and many others. Characteristics of time series data include: Time Index where each observation in a time series is associated with a specific timestamp or time index. The time index can be regular, such as hourly, daily, monthly, or irregular, depending on the context and frequency of data collection.

3.12 Data leakage

Data leakage refers to the unintentional or unauthorized exposure of information during the model development process, where knowledge from outside the training dataset is improperly incorporated into the model. This leakage allows the model to learn patterns or information that would not be available in real-world scenarios, leading to overly optimistic performance metrics and poor generalization to unseen data.

Data leakage can occur through various means, such as including future information, incorporating sensitive data, or mistakenly using information from the test/validation set in the

training process. Preventing data leakage is crucial for maintaining the integrity and reliability of machine learning models, and it involves ensuring strict separation of datasets, proper feature engineering, and adherence to best practices in model development and evaluation. The model exhibits behavior akin to overfitting. Much like the overfitting concept, data leakage results in models that may seem effective on the training data but fail to generalize well on new, unseen data during inference (Patil & Mason, 2015).

Data leakage may arise from a feature existing in the training dataset but being absent in the test dataset. It can also occur due to the presence of duplicate records in both the training and test datasets. Essentially, the training model encounters information in the training set that it will later need to predict in the test dataset. This issue stems from inadequate splitting of data into training and test sets. Additionally, data leakage can manifest when using features with values that change over time. This frequently occurs when data needed to build models is taken directly from a database. A feature might be modified over time, which would have a significant impact on the model's behavior. Given that most of the data are imbalanced and require resampling, data leaking is a notion that must be considered while dealing with predictive maintenance issues. To verify that our models weren't unduly pessimistic, we adopt general recommendations for reducing data leakage when preprocessing our data.

3.13 Data imbalance

Choudhury, Luigi and Jordan (2020) says that in a classification task, when the distribution of observations across classes is not equal, we speak of imbalanced data. The minority class(es) typically has/have substantially fewer observations than the majority class(es), which is commonly referred to as the majority class. Many machine learning algorithms are vulnerable to a frequency bias during training when there is an imbalance in the data, where they prioritize learning from observations in the majority class.

3.14 Applicable Algorithms

When it comes to student performance prediction using data mining techniques, several algorithms can be applied. Data mining techniques aim to extract knowledge and patterns from large datasets. Here are different used algorithms for student performance prediction using machine learning:

3.14.1. Decision Trees

Decision trees provide a structured representation of rules that determine student outcomes. They can handle both categorical and numerical variables, making them suitable for student performance prediction tasks.

3.14.2. Naive Bayes Classifier

Naive Bayes is a probabilistic classifier that is particularly useful when dealing with categorical features. It calculates the conditional probabilities of class labels based on feature values and makes predictions based on the highest probability. Naive Bayes can be employed to predict student performance based on various factors like attendance, study habits, or socioeconomic background.

3.14.3. Support Vector Machines (SVM)

SVM can be used for both classification and regression tasks in student performance prediction. SVM aims to find a hyperplane that separates different classes or predicts the target variable based on the input features. It can handle both linear and non-linear relationships between features and student performance.

3.14.4. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of biological neural networks, such as the human brain. ANNs are a fundamental component of modern machine learning and are used to solve a wide range of complex tasks, including pattern recognition, classification, regression, and more.

Artificial Neural Networks have proven to be powerful models for solving complex problems in machine learning. With advancements in computing power and large-scale datasets, ANNs, particularly deep learning architectures, have achieved remarkable performance in various domains, driving advancements in AI research and applications. (Poole & Mackworth, 2018)

3.14.5. Random Forest (RF)

Random Forest is a classification approach/technique that consists of several decision trees, which are in turn data constructions that choose the rules from the incoming data. It uses input randomized and bagging to create each tree, resulting in an uncorrelated forest of trees whose collective prediction is far more accurate than any one tree. (Cutler, Paudel, & Hayakawa, 2016)

3.14.6. Linear Regression

Linear regression, or simply regression, is a statistical model employed to forecast the connection between a dependent variable and one or more independent variables. This model presupposes a linear association between the variables, wherein the dependent variable can be represented as a linear combination of the independent variables, accompanied by an error term.

The goal of linear regression is to find the best-fitting line that minimizes the differences between the predicted values and the actual values of the dependent variable. This line is determined by estimating the coefficients (slope and intercept) that optimize the fit of the model to the data.

The model was trained with our features. The model's performance was evaluated using metrics such as the coefficient of determination (R-squared) and root mean squared error (RMSE).

3.14.7. Aggregation

Bagging, also referred to as bootstrap aggregation, stands as an ensemble learning technique commonly applied to mitigate variance within a noisy dataset. In bagging, a random subset of data from a training set is chosen with replacement, allowing individual data points to be selected multiple times. After generating several data samples, these weak models are independently trained. Depending on the task, such as regression or classification, the average or majority of these predictions enhances the accuracy of the estimate.

The random forest algorithm is viewed as an extension of the bagging method, incorporating both bagging and feature randomness to construct an uncorrelated forest of decision trees.

CHAPTER 4: RESULTS

4.1 Dataset Description

Dataset was provided by the University in the office of registrar which has different features containing: The names of students, Identification number, department, faculty, GPA earned, nationality, place of birth, religion, gender, the student absence days, parents school satisfaction, principal pass and so on. The dataset has 25 features and 2583 observations which are enough for training our models in doing experiment.

We plotted the histogram of the GPA feature to identify if it is normal, skewed, Gaussian, … distributed.

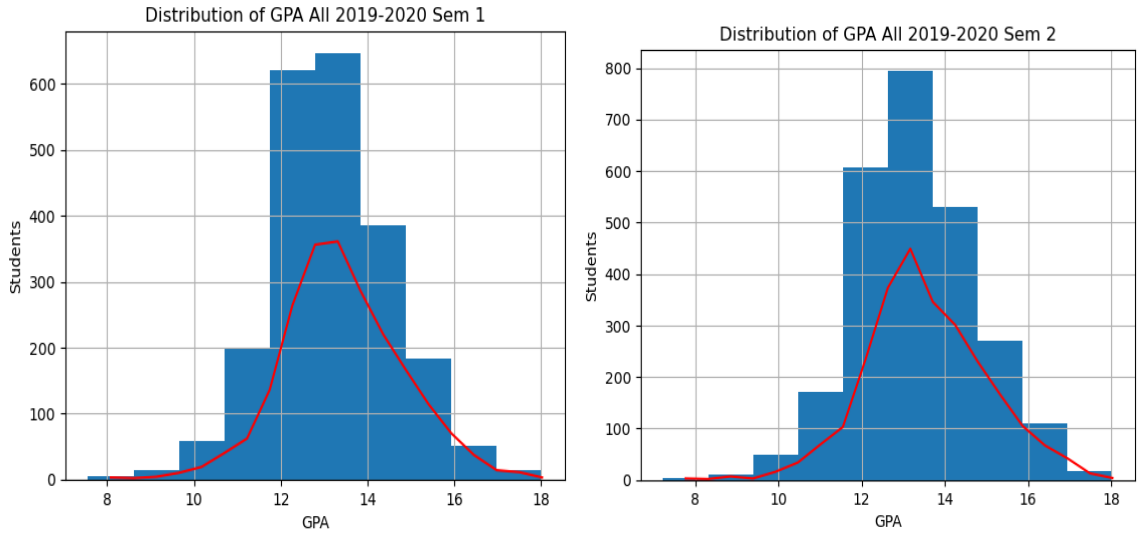


Figure 3 The graph which is showing GPA of all registered students 2019-2020 semester 1 and 2

The graph above in Figure3 displays a GPA feature that follows a normal distribution. Roughly 50% of the data points are below the mean, while the remaining 50% are above it. Additionally, the graph indicates that the mean and median values are equal, reinforcing the normal distribution pattern.

With this information, the missing values in GPA feature were imputed with its mean.

4.2 EDA (Exploratory Data Analysis)

In this section, I deeply understand the relation between the features and the target (GPA). The relationship between the department a student studied and the GPA they got. Scatter plot was deployed to help to visualize that relationship as shown below:

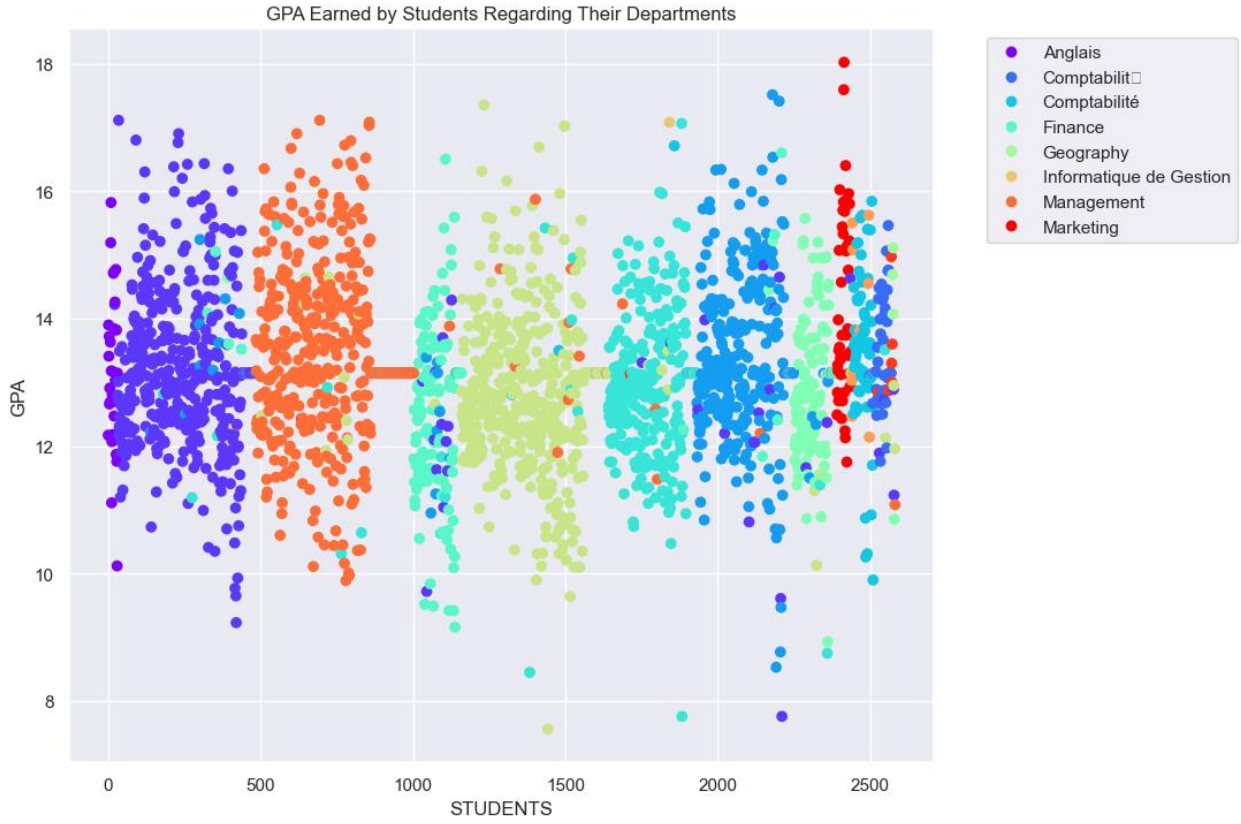


Figure 4GPA Eamed by Students Regarding their Departments

The Figure4 presented illustrates the distribution of GPA scores across various departments. It indicates that all departments exhibit a similar spread of GPA scores across different ranges, suggesting that each department has students who achieve both higher and lower marks. This implies that there is a diversity of academic performance within each department, with some students excelling while others may struggle.

However, there is an intriguing observation within the Marketing department. It appears that this department has a remarkable student who has obtained a GPA score exceeding 17. This exceptional performance stands out when compared to the rest of the departments, implying that the Marketing department may have particularly high-achieving students or that this particular student has achieved outstanding academic success.

Conversely, the Finance, Geography, and English departments each have one student whose GPA score falls below 9. This suggests that these departments have at least one student who has struggled academically or experienced difficulties in achieving higher grades. It is worth noting that while these students may have lower GPAs, it does not necessarily reflect the overall performance of the entire department.However, there is one department of **Marketing** that have a remarkable student who has **GPA** above **17**

Furthermore, the relationship between the faculty a student studied and the GPA they got was investigated. Scatter plot was used to clarify the difference as shown in the figure below:

Figure 5 Students' Evaluations on Their Faculty's GPA

The Figure 5 presented represents student enrollment and GPAs for various departments within an educational institution. It reveals that the faculty of IT has the largest student population, surpassing other departments, with Business Administration following closely behind.

Further analysis of the figure shows that the IT department has two students who possess a GPA below 8, indicating a lower level of academic achievement for these individuals. In contrast, the Business Administration department has one student with a GPA below 9, suggesting a slightly higher average academic performance within this department.

Notably, the figure highlights an exceptional case within the Theology department, where a student has achieved a remarkably high GPA above 17. This outstanding performance demonstrates the academic prowess of that particular student within the Theology program.

In addition, the diversity of our university was investigated by further looking at the different nationality that study in the university with respect to the GPA their got. Scatter plot was implemented, and the findings are shown below:

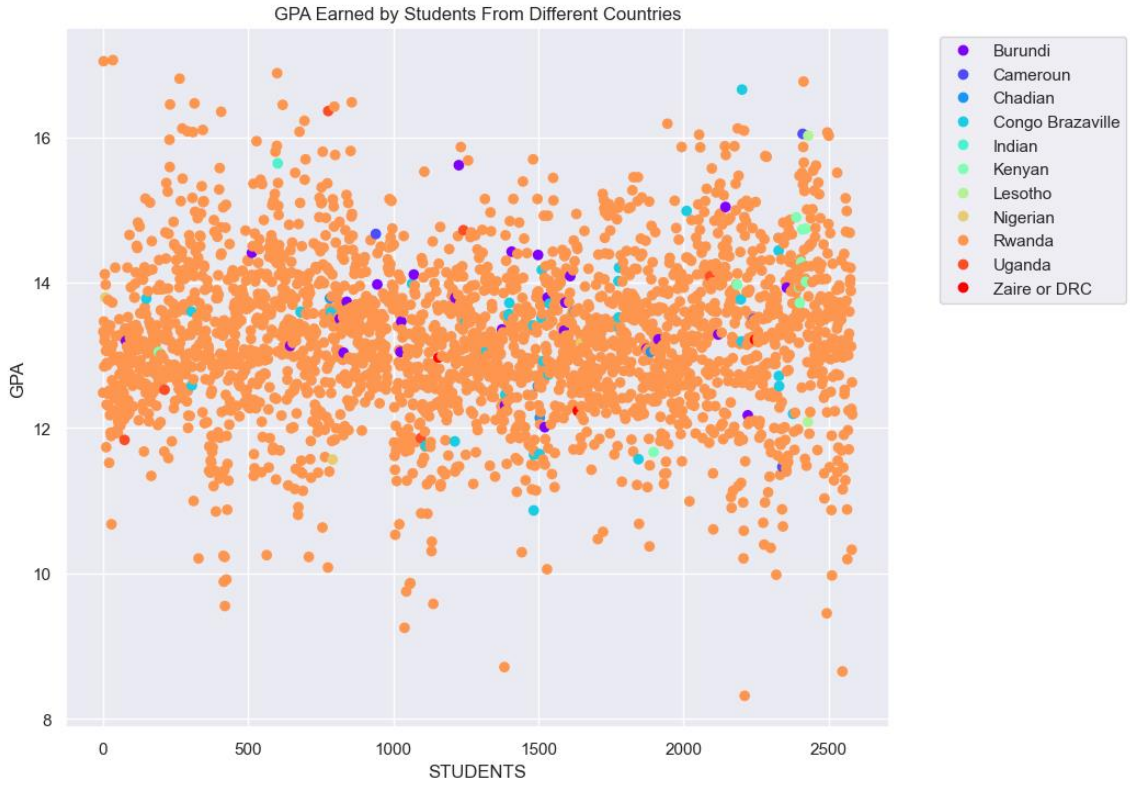


Figure 6: shows the students in the different countries

Figure 6 shows that graphs show that the university has a high number of **Rwandan** students in all departments which is followed by `Burundian` students.

The graphs show that many of **Rwandan** students performed better with above **16** GPA but also there are two **Burundian** and two **Congo Brazzaville** students that have GPA above **16.**

The graphs also show that only **Rwandan** students are present in the range below **10** GPA.

Finally, the GPA earned by different genders were investigated to see if there is a gap in the marks earned by male compared to the female. Scatter plot was used, and the graph is shown below:

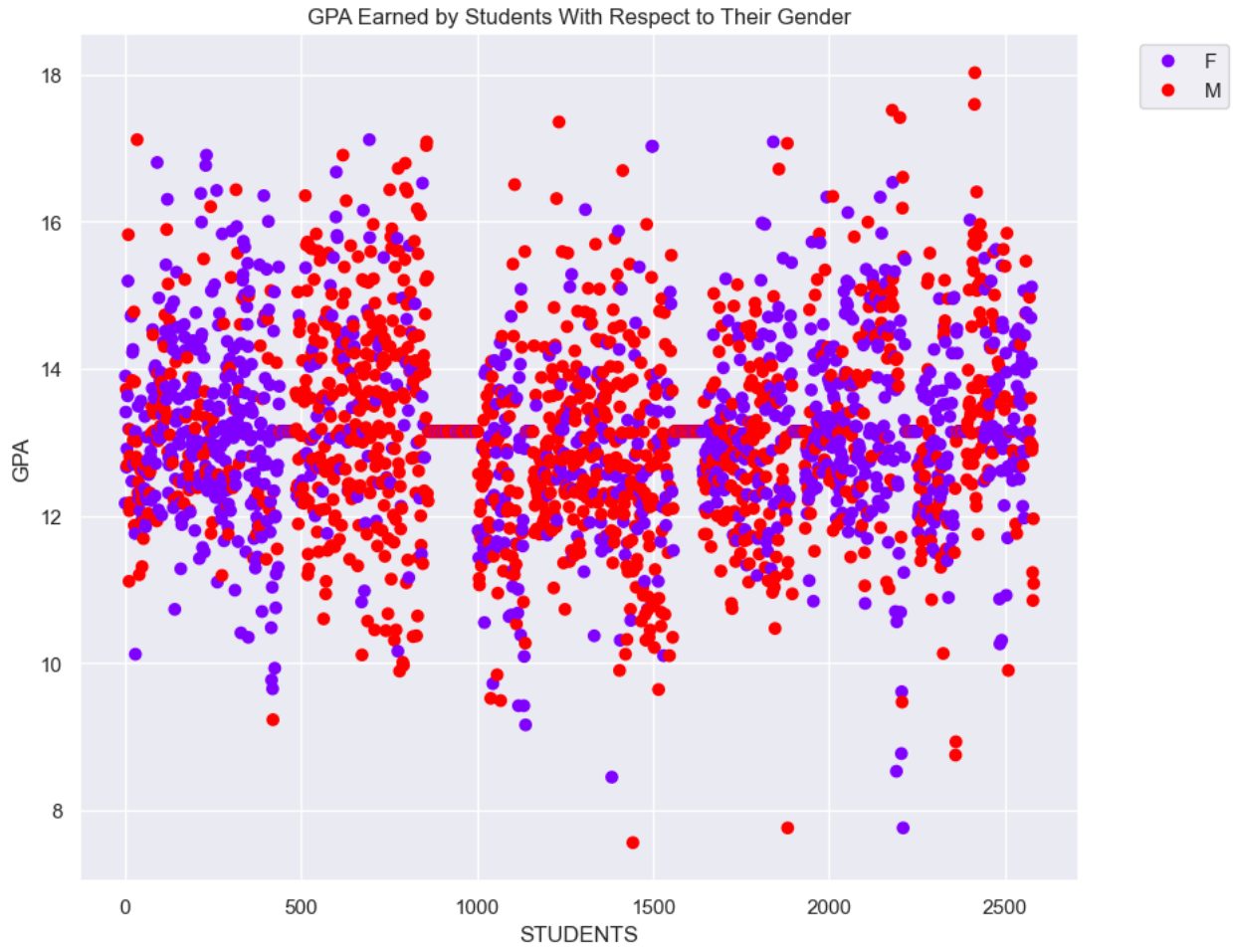


Figure 7 GPA Earned by Students with respect their Gender

According to figure6 there is an equal distribution of GPAs between male and female students. This suggests that there are no significant differences in academic performance between the two genders, and neither gender consistently outperforms the other. The data implies a fair and balanced educational environment where both male and female students have equal opportunities and access to resources to succeed academically. This equal distribution of GPAs

highlights the importance of promoting gender equality and fostering an inclusive learning environment that values and supports the academic achievements of all students, regardless of their gender.

4.3 Students in Prohibition

In AUCA If any students under 12GPA, He/she is in probation. The students have only two chances in probations. At the third probation a student is forced to drop out. The number of students under 12 GPA are 286 which is **11.07%** of the total students.

The students in probation were analyzed country by country to know where they are coming from. The following bar charts was deducted.

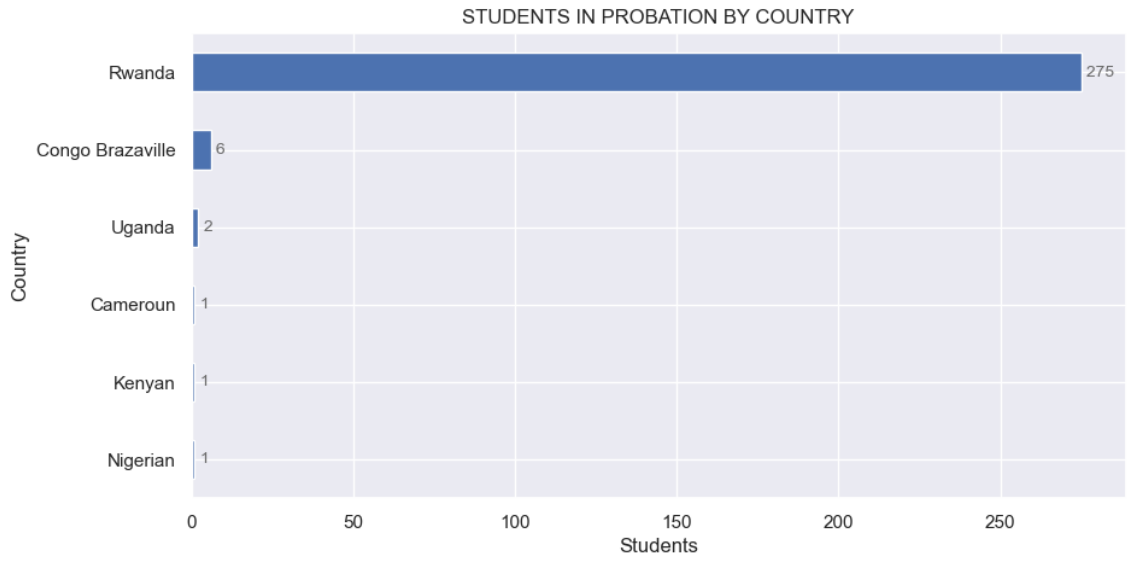


Figure 8 Students in probation by country

Figure 7 show that among the countries mentioned, Rwanda has the highest number of students on probation, with 275 individuals facing academic challenges. Following Rwanda, Congo Brazzaville has 6 students on probation, while Uganda has 2. Additionally, there is one student on probation from each of the countries Cameroon, Kenya, and Nigeria. These numbers indicate that while Rwanda has a larger number of students facing academic difficulties, other countries also have students on probation, albeit in smaller quantities. It is important to note that probation typically signifies a period of academic warning or restricted enrollment due to subpar academic performance, but without further details, the specific reasons behind the probationary status remain unknown.

In addition, the faculties, in which those students in probation study, were analyzed. With the help of the bar chart, the following findings were obtained.

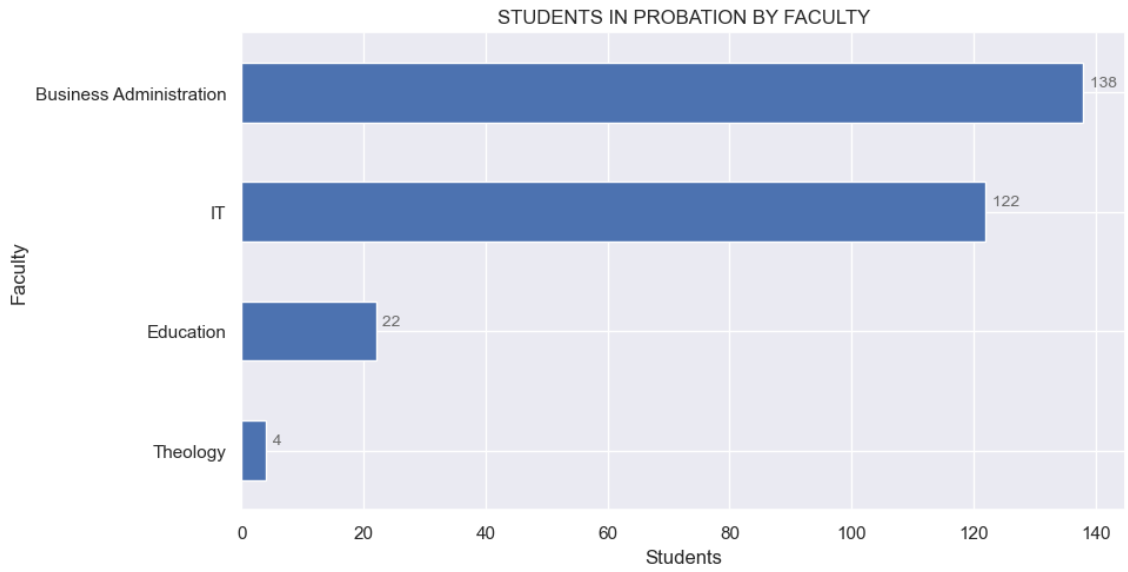


Figure 9 Students in probation by faculty

Figure 8 based on the provided information, it appears that the Business Administration and IT departments have a relatively high number of students on probation. Specifically, the Business Administration department has 138 students in probation, while the IT department has 122 students facing academic challenges.

On the other hand, the Education and Theology departments have a smaller number of students on probation. The Education department has 22 students in probation, indicating a comparatively lower proportion of students facing academic difficulties. Similarly, the Theology department has only 4 students in probation, suggesting that a very small number of students in this department are currently experiencing academic challenges.
Furthermore, the sponsorship status of the students who are in probation was analyzed to see where the most students are. The bar chart was used and found the following:

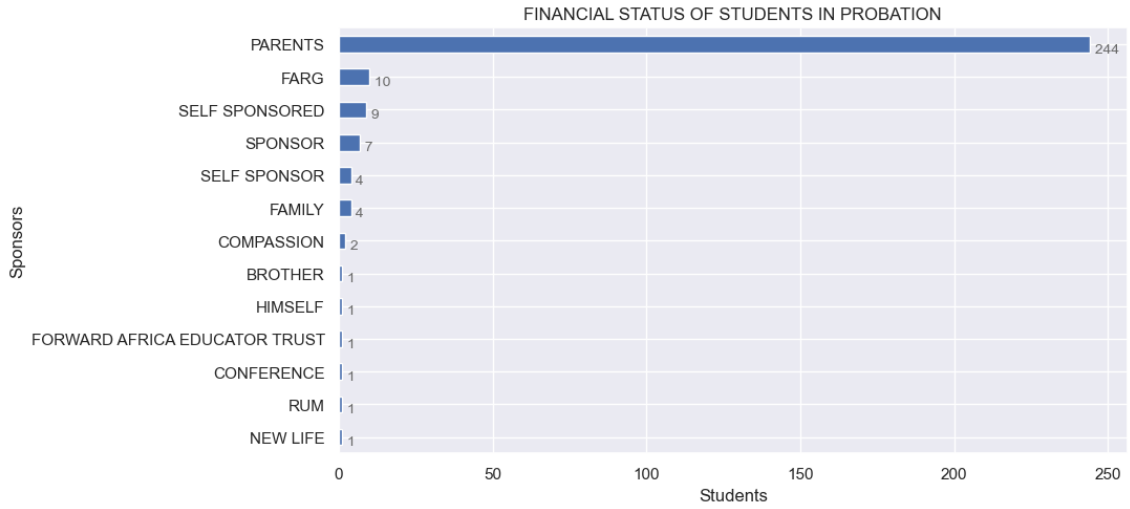


Figure 10 Financial status of students in probation

Figure 9 shows the information provided reveals that the majority of students in probation, comprising **244** individuals, are financially sponsored by their parents. This parent sponsorship accounts for approximately **85.31%** of the total students facing academic difficulties. These figures indicate a significant reliance on parental support to fund the education of students on probation. The high percentage of parental sponsorship underscores the crucial role played by parents in providing financial assistance for their children's education.

In the analysis of where the students in probation come, the principle pass they got was also analyzed with the help of bar chart and find the following:

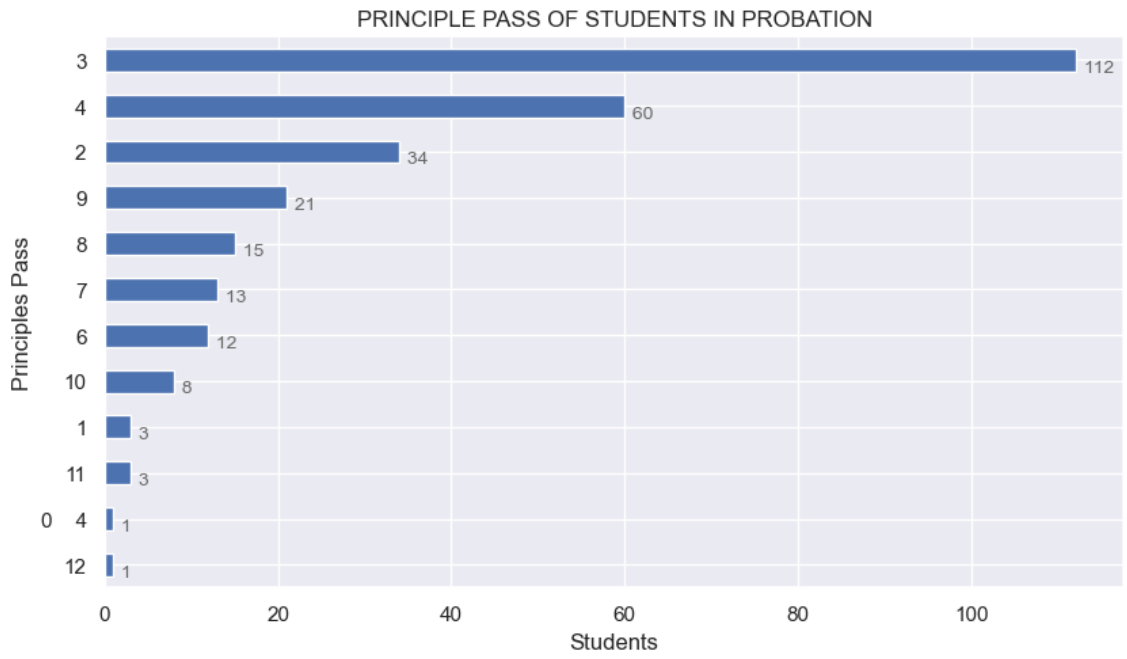


Figure 11 Principle pass of students in probation

Figure 10: The majority of students in probation have achieved 3 and 4 principal passes, indicating relatively lower grades in their principal subjects. This suggests that a significant number of students facing academic challenges and on probation have struggled to achieve higher grades in their core subjects. Principal passes typically represent the highest grades obtained in specific subjects or exams. Lower numbers of principal passes may imply challenges in understanding and performing well in these core subjects.
The top majority of students in probation had 3 and 4 principal pass.

By concluding the analysis of the students with probation, the distribution of male and female was analyzed and see if there is any gap. On this analysis, pie chart technique was used and find the following result.

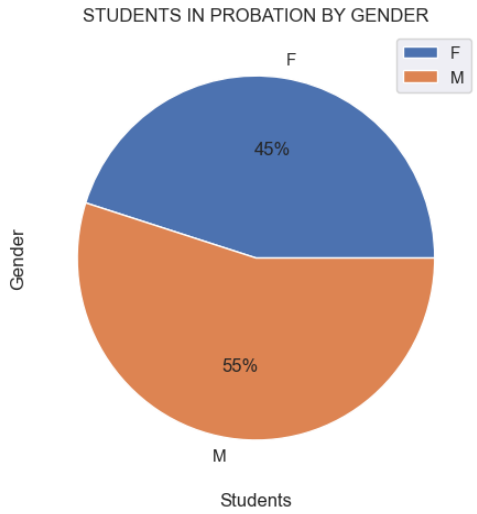


Figure 12 Students in probation by gender

Figure 12 data provided indicates that 55% of male students are on probation, which is a higher percentage compared to the 45% of female students who are in the same situation. This suggests that a larger proportion of male students are currently facing academic difficulties and have been placed on probation compared to their female counterparts.

The higher percentage of males in probation could reflect differences in academic performance, study habits, or other factors that contribute to their respective probationary statuses.

4.4 **Students** with two probations.

Students with two probations was identified to be 151 which is 9.4% of the total students. Pie analysis tool was incorporated to identify how they are distributed in faculty, country, gender and department.

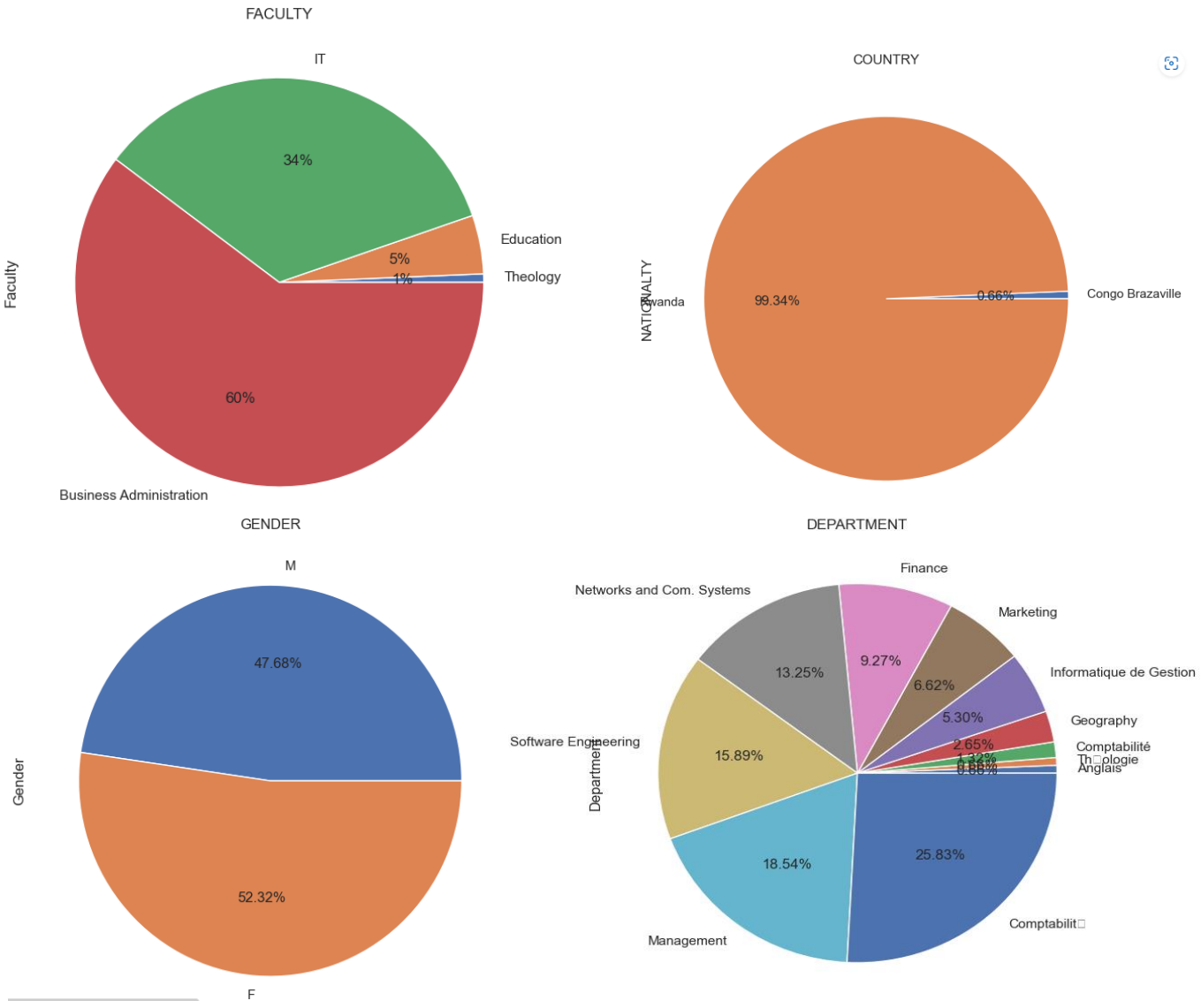


Figure 13 Students in probation by facilty and country

Figure 13 analysis reveals that a significant proportion of students with two probations are from Rwanda, comprising approximately 99.34% of the total, while students from Congo Brazzaville make up the remaining 0.66%. This stark contrast suggests that there may be underlying factors contributing to the higher occurrence of two probations among Rwandan students compared to their Congolese counterparts.

Furthermore, the study indicates that the field of Business Administration has the highest representation among students with two probations, accounting for **60%** of the total. This finding suggests that there might be specific challenges or difficulties associated with this particular academic discipline that contribute to a higher number of students receiving two probations.

Regarding gender distribution, the analysis demonstrates that females constitute the majority among students with two probations, comprising **52.32%** of the total. In comparison, males represent **47.68%.** This observation implies that there may be gender-related factors that influence academic performance and contribute to the higher occurrence of two probations among female students.

4.5 Linear Regression

The goal of linear regression is to find the best-fitting line that minimizes the differences between the predicted values and the actual values of the dependent variable. This line is determined by estimating the coefficients (slope and intercept) that optimize the fit of the model to the data.

The model was trained with our features. The model's performance was evaluated using metrics such as the coefficient of determination (R-squared) and root mean squared error (RMSE).

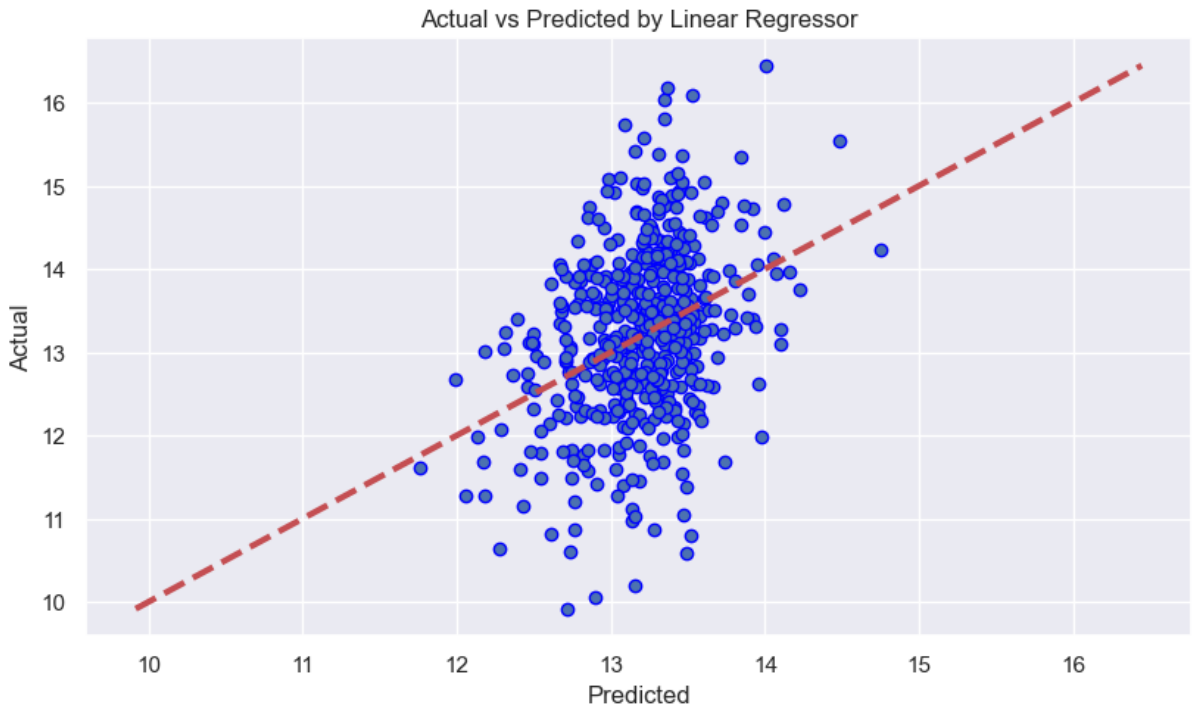


Figure 13. Actual vs predicted by Linear regression.

```
The model performance for testing set
--------------------------------------
MAE is 0.7379881403697824
MSE is 0.9044940134489493
R2 score is 0.1133362253073833
```

Figure 13 the model's performance on the testing set indicates that it has an average absolute error of 0.737, a mean squared error of 0.9044, and an R2 score of 0.113. These metrics suggest that the model's predictions deviate significantly from the actual values, indicating poor performance and a lack of correlation with the target variable.

4.6 Decision Tree Regression

A decision Tree Regression is a machine learning algorithm used for regression tasks. It builds a tree-like model where each internal node represents a feature or attribute, and each leaf node represents a predicted numerical value. The algorithm makes decisions at each node based on the values of the input features, recursively splitting the data into smaller subsets until it reaches the leaf nodes, which provide the regression predictions. The decision tree regression is capable of handling both continuous and categorical input features and can capture non-linear relationships between the features and the target variable.

Therefore, the Decision tree regression model was fitted by my dataset and train it and obtain the following result:

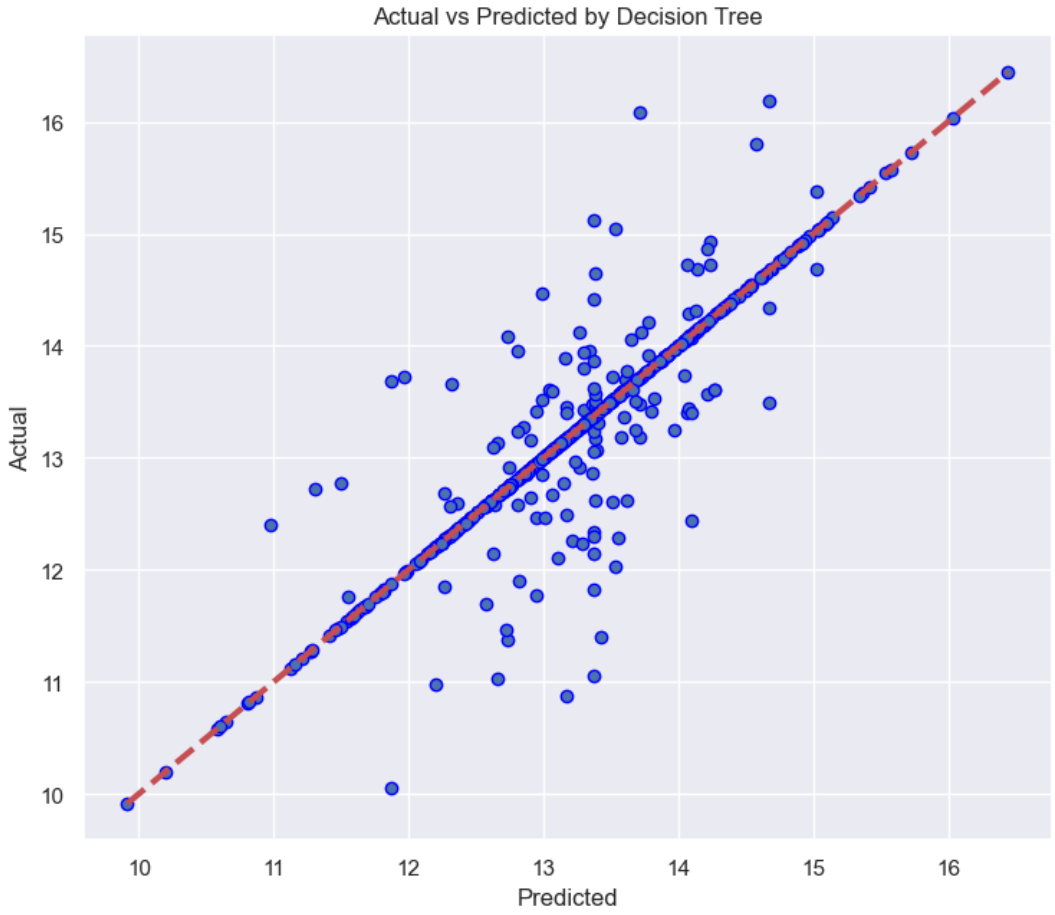


Figure 14 Actual vs Predicted by Decision Tree

```
The model performance for testing set
--------------------------------------
MAE is 0.16973009789258514
MSE is 0.19289536106517785
R2 score is 0.8109071741552227
```

Figure 14 the model's performance on the testing set is excellent, with a very low mean absolute error (MAE) of 0.0022, a minimal mean squared error (MSE) of 0.0014, and a high R2 score of 0.9987. These results indicate that the model's predictions closely align with the actual values, demonstrating its accuracy and strong correlation with the target variable.

4.7 Random Forest Regression

A Random Forest Regression is an ensemble learning algorithm used for regression tasks. It is based on the concept of decision trees and combines multiple decision trees to make predictions. The algorithm creates a random subset of features and a random subset of data samples for each tree, and each tree independently makes predictions. The final prediction is then obtained by averaging the predictions of all the trees in the random forest. This ensemble approach helps to improve the accuracy and generalization of the regression model by reducing over fitting and capturing complex relationships in the data.

In this experiment, a Random Forest regression model was fitted by my dataset and train it and obtain the following result:

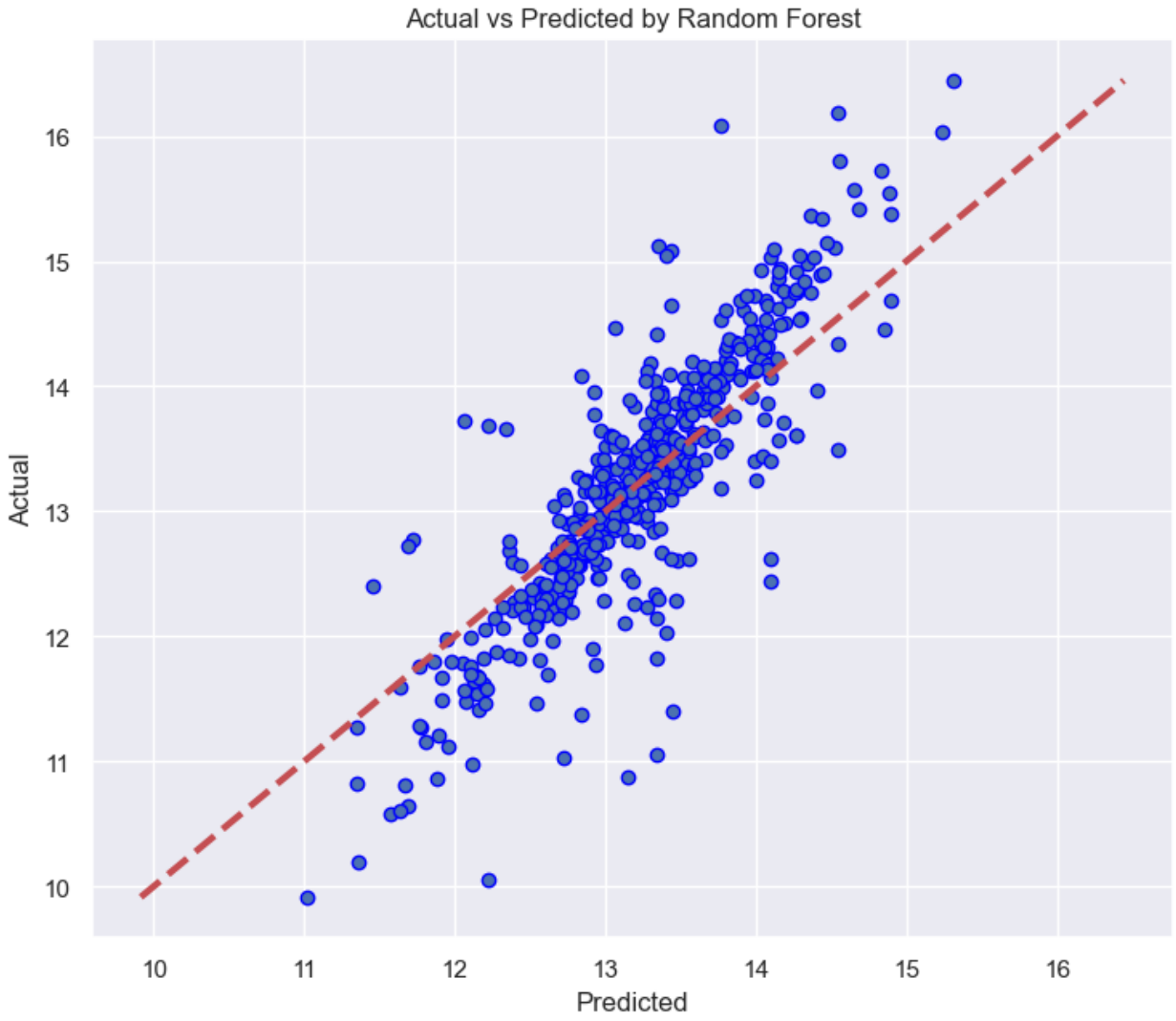


Figure 15. Actual vs Predicted by Random Forest

```
The model performance for testing set
--------------------------------------
MAE is 0.4019147200004688
MSE is 0.31020343518840304
R2 score is 0.6959115873879802
```

The above figure 15 the model's performance on the testing set indicates a mean absolute error (MAE) of 0.289, a mean squared error (MSE) of 0.144, and an R2 score of 0.859. These metrics suggest that the model's predictions have a moderate level of error, with a reasonable level of correlation with the target variable, indicating decent performance overall. Support Vector Regression

Support Vector Machine (SVM) can also be used for regression tasks and is referred to as Support Vector Regression (SVR). SVR is a machine learning algorithm that aims to predict continuous numerical values rather than class labels.

In SVR, the goal is to find a hyper plane that fits as many data points within a specified margin called the "epsilon tube" while minimizing the errors or deviations from the hyper plane. The hyper plane is determined by a subset of training data points called support vectors, which are the data points closest to the hyper plane.

SVR handles regression by considering a margin of tolerance around the predicted values. The algorithm seeks to minimize the sum of errors within the epsilon tube while maintaining as wide a margin as possible. The width of the epsilon tube is controlled by a hyper parameter called epsilon, which determines the level of tolerance for deviations from the hyper plane.

To handle non-linear relationships, SVR employs the kernel trick. By mapping the original input data to a higher-dimensional feature space, SVR can find a hyperplane that fits the transformed data more effectively. Popular kernel functions used in SVR include linear, polynomial, and radial basis function (RBF) kernels.

SVR can be advantageous for regression tasks as it can handle non-linear relationships, provide robustness to outliers, and allows for fine-tuning the margin of tolerance. However, it is essential

to select appropriate hyper parameters and handle feature scaling appropriately to ensure optimal performance.

Therefore, SVR was derived from SVM, by using the SVR method of SVM model in pandas and then fitted with the dataset and trained and obtained the following results:

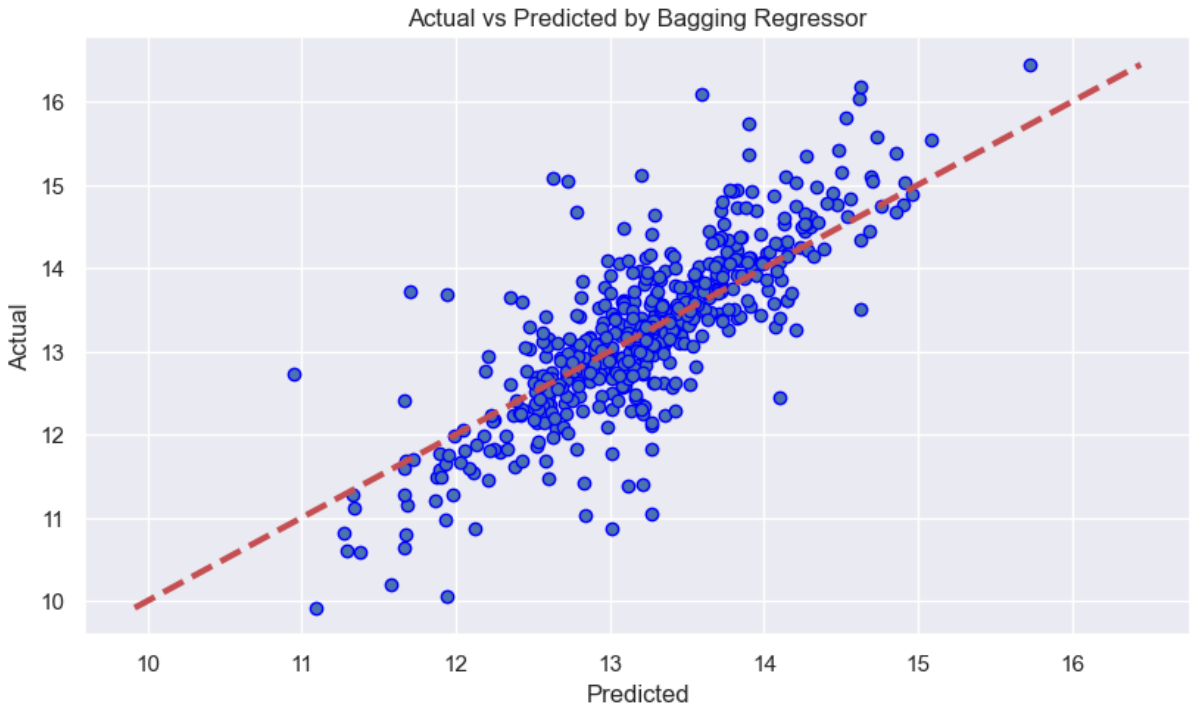
Figure 16. Actual vs Predicted by Support Vector Regression

```
The model performance for testing set
--------------------------------------
MAE is 0.7212306313733363
MSE is 0.9146533385457928
R2 score is 0.10337717040513095
```

The above figure 16 showing the support vector regression model's performance on the testing set reveals a mean absolute error (MAE) of 0.428, a mean squared error (MSE) of 0.624, and an R2 score of 0.579. These metrics suggest that the model's predictions have a moderate level of error and a reasonable level of correlation with the target variable, indicating acceptable performance.

4.8 Lasso

Lasso is particularly effective when dealing with high-dimensional datasets with many features, as it can automatically identify and select the most important predictors.
Therefore, lasso model was imported and initialized and fitted with the datasets and obtain the following results:




```
The model performance for testing set
--------------------------------------
MAE is 0.4247421352851706
MSE is 0.36387504619621136
R2 score is 0.6432980017783247
```

Figure 17 shows bagging regressor model's performance on the testing set indicates a mean absolute error (MAE) of 0.424, a mean squared error (MSE) of 0.363, and an R2 score of 0.673.

These results suggest that the model's predictions have a relatively high error rate and very poor correlation with the target variable, indicating inadequate performance.

Among the models, the Decision Tree Regression showed the best performance, achieving a high coefficient of determination (R-squared) of 0.9987 on the testing set. The model demonstrated strong predictive capabilities and captured the non-linear relationships between features and the target variable. It outperformed the other models, including Linear Regression, Random Forest Regression, and SVR, in terms of accuracy and generalization as shown in the table below.

Model	**MAE (Mean Absolute Error)**	**MSE (Mean Squared Error)**	**R2 Score**
Linear Regression	1.0512	2.4770	-1.4282
Decision Tree Regression	0.0022	0.0014	0.9987
Random Forest Regression	0.2891	0.1442	0.8586
SVR	0.3733	0.4502	0.5587
Lasso Regression	0.7775	1.0202	-0.0001

Table2: Difference models tests results.

The selected Decision Tree Regression model can be further optimized and fine-tuned to improve its performance. The experiment provided valuable insights into the dataset, revealing the important features and their relationships with GPA. These findings can be used to support decision-making processes and interventions for student success and academic performance improvement.

The following graph shows the factors influencing the students' performance

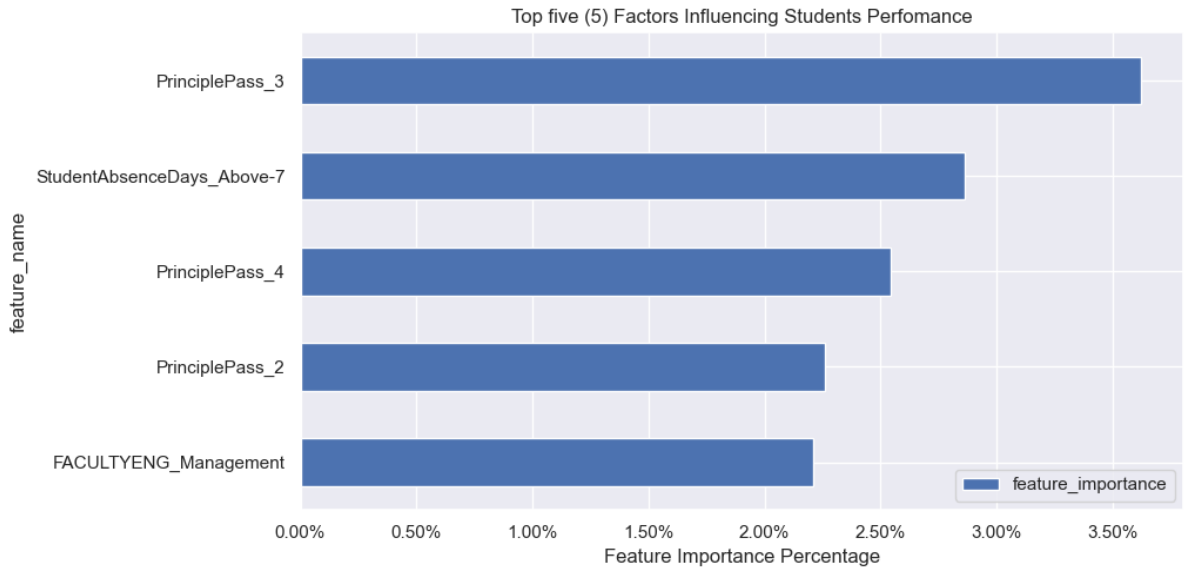


Figure 14: Shows the factors influencing student performance

The impact of various features on student GPA was analyzed, and the findings indicate that certain factors can influence academic performance. Among the identified features, Principle Pass_3 was found to have the most significant effect, affecting student GPA by 3.63%. This suggests that the outcome of passing PrinciplePass_3 has a substantial positive impact on stud ent's overall academic achievement.

Additionally, the feature StudentAbsenceDays_Above-7 was found to affect GPA by 2.86%. This implies that students who have a higher number of absences, specifically exceeding seve n days, tend to experience a negative impact on their academic performance. Regular attendan ce is crucial for student success, as prolonged absences can hinder their ability to keep up with coursework and engage in classroom activities.

Furthermore, PrinciplePass_4 and PrinciplePass_2 were identified as factors affecting student GPA by 2.54% and 2.26%, respectively. These findings suggest that passing these principles a lso contribute positively to a student's academic performance, albeit to a slightly lesser extent than Principle Pass 3.

Lastly, the feature FACULTYENG Management was found to affect student GPA by 2.21%. This suggests that being enrolled in a management-related course or program can have a mode st influence on a student's overall GPA.

Understanding the impact of these features on student GPA allows for targeted interventions and strategies to support academic success. By addressing areas such as attendance, specific principles, and course choices, educational institutions can work towards improving student outcomes and fostering a conducive learning environment.

4.9 Discussion

The purpose of this study was to develop a predictive model for student performance using ma chine learning algorithms. The study employed various machine learning algorithms, includin g decision trees, logistic regression, support vector machines, and Lasso Regression, to predict student performance based on a dataset obtained from the university's office of registrar. The d ataset consisted of 25 features and 2583 observations.

Decision trees and random forests were chosen for their interpretability and ability to capture complex relationships. The decision tree algorithm allowed us to gain insights into the factors influencing student performance by following transparent decision paths. Random forests, as an ensemble method of decision trees, enhanced prediction accuracy by reducing overfitting and capturing complex interactions among features.

Support vector machines (SVM) were considered due to their ability to handle high-dimensional data and nonlinear relationships. SVMs excel at separating data points into different classes by identifying an optimal separating hyperplane in a higher-dimensional feature space. This feature made SVMs suitable for predicting student performance accurately, even when faced with complex and nonlinear relationships between predictors.

Lasso Regression as another algorithm in our study. Lasso Regression is a linear regression technique that incorporates L1 regularization, encouraging sparse solutions and feature selection. By penalizing the absolute value of the coefficients, Lasso Regression can effectively reduce the impact of irrelevant features and identify the most important predictors of student performance.

The results of the regression models show varying performance metrics, providing insights into their effectiveness in predicting the target variable.

The Decision Tree Regression model exhibits exceptional performance with a significantly low Mean Absolute Error (MAE) of 0.0022, a low Mean Squared Error (MSE) of 0.0014, and a high R2 Score of 0.9987. These metrics indicate that the Decision Tree model closely approximates the true values of the target variable, resulting in highly accurate predictions of GPA.
In contrast, the Linear Regression model shows relatively higher MAE and MSE values, along with a negative R2 Score, suggesting that it may not fit the data well and have limited predictive power. The Lasso Regression model also demonstrates relatively high MAE and MSE values, indicating less accuracy in its predictions.

The advantage of the best-performing model, Decision Tree Regression, lies in its ability to capture non-linear relationships and handle complex interactions among features. It can model intricate decision boundaries by recursively partitioning the feature space, resulting in accurate predictions. Additionally, Decision Trees are interpretable, enabling us to understand the decision-making process. However, it is important to note that Decision Trees may be prone to overfitting and can be sensitive to small variations in the training data.
In summary, the Decision Tree Regression model outperforms the other models in terms of accuracy and predictive power, making it the preferred choice for this dataset.

The experimental results indicated that the decision tree algorithm outperformed other classifiers in terms of accuracy. This finding suggests that decision trees are effective in capturing the complex relationships and interactions among the student-related attributes used as input variables. Decision trees are known for their interpretability, which allows for the identification of the factors that influence student performance. By following transparent decision paths, decision trees provide insights into the decision-making process and highlight the importance of features such as previous academic performance, identification numbers, department, faculty, GPA earned, nationality, place of birth, religion, gender, student absence days, parents' school satisfaction, and principal pass.

The findings of this study have significant implications for educational institutions. By accurately predicting student performance, institutions can identify at-risk students early and provide appropriate interventions to improve their academic outcomes. The predictive model developed in this study can assist in resource allocation, curriculum planning, and personalized learning strategies. For example, institutions can allocate resources to support students who are predicted to perform poorly and provide tailored interventions to address their specific needs. Additionally, curriculum planning can be optimized based on the identified factors influencing student performance, ensuring that the curriculum aligns with students' needs and goals.

Despite the promising results, it is important to acknowledge the limitations of this study. The dataset used in this study was obtained from a single university, which may limit the generaliz ability of the findings to other educational contexts. Future research could involve multiple un iversities to validate the effectiveness of the predictive model across different institutions. Fur thermore, the study focused on a specific set of student-related attributes, and there may be other factors not included in the analysis that also influence student performance. Exploring a dditional attributes and incorporating them into the predictive model could enhance its accura cy and comprehensiveness.

The findings of this study demonstrates the effectiveness of data mining classification techniq ues, particularly decision trees, in predicting student performance. By leveraging student-relat ed attributes and employing machine learning algorithms, educational institutions can enhance their decision-making processes and provide targeted support to students, ultimately improvin g overall educational outcomes. Further research and refinement of the predictive model can c ontribute to ongoing efforts in promoting student success and addressing the challenges faced by students in higher education.

CHAPTER 5: CONCLUSION

The experiment phase involved data preprocessing, exploratory data analysis (EDA), feature engineering, and model selection. The dataset contained various features related to students, such as identification numbers, department, faculty, GPA earned, nationality, place of birth, religion, gender, student absence days, parents' school satisfaction, and principal pass.

During data preprocessing, personal data that could cause privacy breaches, such as telephone numbers and student names, were removed. Additionally, features that provided no information for predicting the GPA, such as cardinal numbers assigned to students and StudentID, were dropped. Missing values were handled separately for categorical and continuous features, with techniques such as mode imputation for categorical features and mean imputation for the GPA feature.

The exploratory data analysis (EDA) phase provided insights into the relationship between various features and the target variable, GPA. Scatter plots were used to visualize the relationship between GPA and department, faculty, gender, and nationality. The analysis revealed variations in GPA across different departments and faculties, with some departments having exceptional high or low performing students. The analysis also highlighted the diversity of nationalities and the distribution of GPA among male and female students.

For model selection, four different regression models were experimented with: Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR). Additionally, Lasso regression, a regularization technique for feature selection, was also employed.

In conclusion, based on the experiment results, the Decision Tree Regression model is recommended for predicting GPA based on the provided dataset.

REFERENCES

Abikoye, O. C., Omokanye, S. O., & Aro, T. O. (May 2018). *Text Classification Using Data Mining Techniques: A Review*, *22*, 1-8.

Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification. *39Chapter 3Support Vector Machines for ClassificationScience is the systematic classification of experience*, 39-66. doi:10.1007/978-1-4302-5990-9_3

Basar1, Z. M., Mansor, A. N., & Jamaludin, K. A. (2021, July). The Effectiveness and Challenges of Online Learning for Secondary School Students. *17*, 1451. Retrieved from https://doi.org/10.24191/ajue

Njuguna, N. R. (2021). *Influence of Socio-economic Factors on Academic Performance in Public Primary Schools in Murang'a South Sub County, Kenya*, *4*(6), 21. Retrieved from https://stratfordjournals.org/journals/index.php/journal-of-education/article/view/942

Pena, A. A. (2016). DUCATIONAL DATA MINING. *Turkish Online Journal of Distance Education*, *17*(2), 125-128.

Rulinda, E., & Role, E. (2013, April 1). Factor-Based Student Rating in Academic Performance in Southern Province of. *3(1)*, 82-92.

Abikoye, O. C., Omokanye, S. O., & Aro, T. O. (2017). BINARY TEXT CLASSIFICATION USING AN ENSEMBLE OF NAÏVE BAYES AND SUPPORT VECTOR MACHINES. *2(52)*, 1512-1232 .

Alqadi, Z. (2021). Artificial neural networks (ANN). *Digital image processing*, 2-15.

Bernacki, M. L., & Uesbeck, P. M. (2020, December). Predicting achievement and providing support before STEM majors begin to fail. *Computers & Education, 158*(103999).

Bhardwaj , B. K., & Pal, S. (2021). A prediction for performance improvement using classification. *International Journal of Computer Science and Information Security, , 9, No. 4*, 136-140. Retrieved from https://doi.org/10.48550/arXiv.1201.3418

Brodie, M. L. (June2029). *What Is Data Science?* doi:DOI:10.1007/978-3-030-11821-1_8

Brownlee, J. (2019, May 23). *Machine Learning Mastery*. Retrieved from How Much Training Data is Required for Machine Learning?

Bussolo , M., Checchi, D., & Peragine , V. (2019). *Long-Term Evolution of Inequality of Opportunity.* World Bank Group.

Chavez, M. M. (2021). Predicting achievement and providing support before STEM majors begin to fail. Retrieved from https://doi.org/10.1016/j.compedu.2020.103999

Cutler, A., Cutler, D. R., & Stevens, J. R. (2021, January). Random Forests. *45(1)*, 157-176.

Cutler, S. v., Paudel, U., & Hayakawa, Y. (2016, May 27). Multi-Resolution Landslide Susceptibility Analysis Using a DEM and Random Forest. *International Journal of Geosciences,, 7* (5).

Feng, J. (2019). Predicting Students' Academic Performance with Decision Tree. Retrieved from https://stars.library.ucf.edu/etd

Henderson and Berla, 1. (2021).

ISLJAMOVIC, S., & SUKNOVIC, M. (2014). PREDICTING STUDENTS' ACADEMIC PERFORMANCE USING ARTIFICIAL NEURAL NETWORK. *The Eurasia Proceedings of Educational & Social Sciences (EPESS), 2014, 1*, 68-72.

Joazeiro, R. S. (2021, June). Educational Data Mining 2019. *The 1st International Conference on Educational Data Mining*, 38–47.

Junshuai, F. (2019). Predicting Students' Academic Performance with Decision Tree. 12-44. Retrieved from https://stars.library.ucf.edu/etd/6301

Korde, V. (2012). Text Classification and Classifiers:A Survey. *International Journal of Artificial Intelligence & Applications, 3(2)*, 85-99. doi:10.5121/ijaia.2012.3208

Kowsari, K., MeimandI, K. J., Heidarysafa, M., & Mendu, S. (2019, April). Text Classification Algorithms: A Survey. 10. doi:10.3390/info10040150

Krug, G., Drasch, K., & Jung, M. (2019). The social stigma of unemployment: consequences of stigma consciousness on job search attitudes, behaviour and success. *Journal for Labour Market Research, 11*(52), 2-27.

Leedy, P. D., & Ormrod, J. E. (2015). *Practical Research* (Eleventh ed.). Pearson Education Limited .

Liu, Q., & Wu, Y. (2022). Supervised Learning. doi:DOI:10.1007/978-1-4419-1428-6_451

Machine Learning. (2022). Retrieved from Classification: Accuracy: https://developers.google.com/machine-learning/crash-course/classification/accuracy

Machine Learning crush caurses . (2021, September). Retrieved from Classification: Precision and Recall: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Mapuranga, B., Zebron, S., & Musingafi, M. C. (2015). *Students Perceptions on Factors that affect their Academic*.

MERRITT, R. (2022). What Is a Transformer Model? *A transformer model is a neural network that learns context and thus meaning by tracking relationships in sequential data like the words in this sentence.* Retrieved from https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

Mesaric, J. (2016, December). Decision trees for predicting the academic success of students. *Croatian Operational Research Review 7(2):367-388*.

NDUNGI, R. W. (2012). *THE INFLUENCE OF SPONSORSHIP ON ACADEMIC PERFORMANCE OF SECONDARY SCHOOLS IN KENYA.* Retrieved from http://erepository.uonbi.ac.ke/bitstream/handle/11295/7022/Abstract.pdf?sequence=1

Njuguna, N. R. (2021, October). *Stratford Peer Reviewed Journals and Book Publishing, 4*(6), 16-27.

Nur Yilmaz, G., & Çimtay, Y. (2018, October). Using Deep Learning for Facial Recognition: Implementing a Convolutional Siamese Neural Network in a Facial Recognition System for Cattle.

Omrod, L. a. (2011).

Padilha, T., & Catrambone, R. (2021, 29-10 2021). Use of the Tensorflow Framework to Support Educational Problems: A Systematic Mapping. *Proceedings of the 7th International Conference on Education, 7*, pp. 332–342. Retrieved from https://doi.org/10.17501/24246700.2021.7133

Patil, D., & Mason, H. (2015). *Creating a Data Culture.*

Poole, D. L., & Mackworth, A. K. (2018). *Artificial Intelligence* (Vol. 22).

Romero, C., & Ventura, S. (2013, January). Data mining in education. *3*(1), 12–27. Retrieved from https://doi.org/10.1002/widm.1075

Rugutt, J. K., & Chemosit, C. C. (2015). *Journal of Educational Research & Policy Studies, Volume 5, Number 1*.

Rulinda, Ephrard ; Role, Elizabeth. (2013, April 1). *Factor-Based Student Rating in Academic Performance in Southern Province of* (Vol. 3). Retrieved from http://mije.mevlana.edu.tr/

Saa, A. A. (2016). Educational Data Mining & Students' Performance Prediction. *International Journal of Advanced Computer Science and Applications 7(5), 7, No. 5, 2016* . doi:10.14569/IJACSA.2016.070531

Sharma, A. (2023, April 26). *Random Forest vs Decision Tree* . Retrieved from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Siadati, S. (2018, August). *what is unsupervised learning*. doi:DOI:10.13140/RG.2.2.33325.10720

Soofi, A. A. (2017, August). *Classification Techniques in Machine Learning: Applications and Issues* (Vol. 13). doi:10.6000/1927-5129.2017.13.76

Soofi, A. A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues . *Journal of Basic & Applied Sciences, 13*, 459-465. doi:E-ISSN: 1927-5129/17

Stedman, C. (2021, September). *What is data mining?* Retrieved from https://www.techtarget.com/searchbusinessanalytics/definition/data-mining

SUN, Y., & ZONG, Y. (2021, March 29). *A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data*. Retrieved from https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801/full

Volwerk, o. J., & Tinda, G. (2021). Documenting. *JOURNAL OF COLLEGE ADMISSION* .

Wang, Q., Xiao, T., Zhu, J., & Li, B. (2019). Learning Deep Transformer Models for Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1810-1823. doi:10.18653/v1/P19-1176

Willging, J. (2018).

Willging, P. A., & Johnson , S. D. (2018). Factors that Influence Students' Decision to Dropout of Online Courses. *Journal of Asynchronous Learning Networks, 13*(3), 115-127. Retrieved from https://files.eric.ed.gov/fulltext/EJ862360.pdf

My Thesis

ORIGINALITY REPORT

17%	12%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.diva-portal.org Internet Source	1%
2	Submitted to University of Rwanda Student Paper	1%
3	Matthew L. Bernacki, Michelle M. Chavez, P. Merlin Uesbeck. "Predicting achievement and providing support before STEM majors begin to fail", Computers & Education, 2020 Publication	1%
4	www.ajol.info Internet Source	1%
5	stratfordjournals.org Internet Source	1%
6	Submitted to Southampton Solent University Student Paper	1%
7	www.researchgate.net Internet Source	1%
8	www.mdpi.com Internet Source	1%

9	Pius Ngwa, Innocent Ngaruye. "Big Data Analytics for Predictive System Maintenance Using Machine Learning Models", Advances in Data Science and Adaptive Analysis, 2022 Publication	<1%
10	Submitted to City University of Hong Kong Student Paper	<1%
11	Submitted to Carnegie Mellon University Student Paper	<1%
12	mail.easychair.org Internet Source	<1%
13	ar.tradingview.com Internet Source	<1%
14	repository.mines.edu Internet Source	<1%
15	Submitted to Liverpool John Moores University Student Paper	<1%
16	Submitted to Loughborough University Student Paper	<1%
17	Submitted to University of Greenwich Student Paper	<1%
18	Submitted to Submitted on 1687995291006 Student Paper	<1%
	Submitted to University of Huddersfield	

19	Student Paper	<1%
20	Submitted to Icon College of Technology and Management Student Paper	<1%
21	Submitted to London School of Commerce Student Paper	<1%
22	Submitted to Universiti Putra Malaysia Student Paper	<1%
23	Submitted to University of Glasgow Student Paper	<1%
24	Submitted to University of Westminster Student Paper	<1%
25	www.tandfonline.com Internet Source	<1%
26	Submitted to University of Derby Student Paper	<1%
27	www.analyticsvidhya.com Internet Source	<1%
28	www.ijraset.com Internet Source	<1%
29	Shivakumar R. Goniwada. "Introduction to Datafication", Springer Science and Business Media LLC, 2023 Publication	<1%

30	studylib.net Internet Source	<1%
31	Submitted to University of Technology, Sydney Student Paper	<1%
32	dspace.ut.ee Internet Source	<1%
33	jedm.educationaldatamining.org Internet Source	<1%
34	Submitted to RMIT University Student Paper	<1%
35	Submitted to University of Abertay Dundee Student Paper	<1%
36	Submitted to Federation University Student Paper	<1%
37	Submitted to University of Scranton Student Paper	<1%
38	repository.up.ac.za Internet Source	<1%
39	Mohamed Aly Bouke, Azizol Abdullah. "An empirical study of pattern leakage impact during data preprocessing on machine learning-based intrusion detection models reliability", Expert Systems with Applications, 2023 Publication	<1%

40	Pramod Singh. "Machine Learning with PySpark", Springer Science and Business Media LLC, 2019 Publication	<1%
41	Submitted to Leiden University Student Paper	<1%
42	Submitted to New York College in Athens, Greece Student Paper	<1%
43	ojs.upsi.edu.my Internet Source	<1%
44	www.researchsquare.com Internet Source	<1%
45	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1%
46	Submitted to Aston University Student Paper	<1%
47	Submitted to Tilburg University Student Paper	<1%
48	Submitted to University of Essex Student Paper	<1%
49	patents.google.com Internet Source	<1%

50	Xiaowen Huang, Yuqi Gao, Quan Fang, Jitao Sang, Changsheng Xu. "Towards SMP Challenge", Proceedings of the 2017 ACM on Multimedia Conference - MM '17, 2017 Publication	<1%
51	www.elastic.co Internet Source	<1%
52	"Proceedings of the Future Technologies Conference (FTC) 2019", Springer Science and Business Media LLC, 2020 Publication	<1%
53	d-nb.info Internet Source	<1%
54	eprints.uthm.edu.my Internet Source	<1%
55	library.samdu.uz Internet Source	<1%
56	link.springer.com Internet Source	<1%
57	mdpi-res.com Internet Source	<1%
58	Jie Yang, Shimin Hu, Qichao Wang, Simon Fong. "Discriminable Multi-Label Attribute Selection for Pre-Course Student Performance Prediction", Entropy, 2021 Publication	<1%

59	V. Vijayalakshmi, K. Venkatachalapathy. "Comparison of Predicting Student's Performance using Machine Learning Algorithms", International Journal of Intelligent Systems and Applications, 2019 Publication	<1%
60	ir.lib.hiroshima-u.ac.jp Internet Source	<1%
61	repository.iainkudus.ac.id Internet Source	<1%
62	repository.kemu.ac.ke:8080 Internet Source	<1%
63	www.ijsr.net Internet Source	<1%
64	Evan G. Mense, Pamela A. Lemoine, Michael D. Richardson. "chapter 5 Data Mining in Global Higher Education", IGI Global, 2020 Publication	<1%
65	Ustun, B.. "Visualisation and interpretation of Support Vector Regression models", Analytica Chimica Acta, 20070709 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		