Design And Implementation of the BAZE University Student Dropout Prediction System Model using Machine learning Algorithms

Thesis/Report submitted in partial fulfilment of the requirement

for the degree of

B.Sc.

In

Information Systems Management

By

Mustapha Adam

To

The Department of Computer Science

Baze University, Abuja

[Month Year]

# DECLARATION

This is to certify that this Thesis/Report entitled Design and Implementation of BAZE University Student Dropout Prediction system machine learning model, which is submitted by Mustapha Adam in partial fulfillment of the requirement for the award of degree for B.Sc. in Information Technology to the Department of Computer Science, Baze University Abuja, Nigeria, comprises of only my original work and due acknowledgement has been made in the text to all other materials used.

Date: [Date Month Year] Name of Student: [Your Name]

**APPROVED BY** …………………

**Dean**

Faculty of Computing and Applied Sciences

# CERTIFICATION

This is to certify that this Thesis/Report entitled Design and Implementation of BAZE University Student Dropout Prediction system machine learning model, which is submitted by Mustapha Adam in partial fulfillment of the requirement for the award of degree for B.Sc. in Information Technology to the Department of Computer Science, Baze University Abuja, Nigeria is a record of the candidate’s own work carried out by the candidate under my/our supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Date: Supervisor:

# APPROVAL

This is to certify that the research work, Design and Implementation of BAZE University Student Dropout Prediction System machine learning model and the subsequent preparation by Mustapha Adam with [Student ID] has been approved by the Department of Computer Science, Faculty of Computing and Applied Science, Baze University, Abuja, Nigeria.

By

[Full name] Date

1st Supervisor

[Full name] Date

2nd Supervisor

Dr. Moses Ubaru Date

Head of Department

Prof Peter Ogedebe Date

Dean, Faculty of Computing and Applied Science

[Full name] Date

External Examiner

# DEDICATION

[This is the dedication page.]

# ABSTRACT

*High Institutions face significant challenges related to student attrition and dropout rates. These challenges not only impact the students' educational journeys but also have financial and operational implications for universities. To address this issue, the "Baze University Student Dropout Prediction System Model" has been designed and implemented. This system leverages machine learning algorithms to predict students at risk of dropping out, allowing for timely interventions and support. The system begins by collecting and integrating diverse student data, including academic records, socio-demographic information, and historical dropout data. The data undergoes pre-processing, including handling missing values and outliers, to ensure data quality. Feature selection and engineering techniques are employed to identify the most relevant predictors of dropout. Three machine learning algorithms—Random Forest, Decision Tree, and Logistic Regression—are made available for dropout prediction. Users can select and train these models on historical data, with the option for hyperparameter tuning to optimize performance. Real-time data updates ensure that the model remains adaptive to changing student profiles.*

*It offers a range of accuracy and performance metrics, such as accuracy, precision, recall, and F1-score, to evaluate the model's reliability. Visualization tools and reporting capabilities enable stakeholders to comprehend results effectively. Furthermore, the system integrates with student support services, facilitating early interventions for identified at-risk students. This proactive approach enhances student success rates and promotes a supportive learning environment.*

*The "BAZE University Student Dropout Prediction System Model" represents a crucial tool for educational institutions, empowering them to reduce dropout rates, allocate resources more effectively, and foster student success, it is recommended that the model developed will be deployment on mobile and web application.*

Table of Contents

[DECLARATION ii](#_Toc145753842)

[CERTIFICATION iii](#_Toc145753843)

[APPROVAL iv](#_Toc145753844)

[DEDICATION v](#_Toc145753845)

[ABSTRACT vi](#_Toc145753846)

[CHAPTER ONE INTRODUCTION 1](#_Toc145753847)

[1.1 Overview 1](#_Toc145753848)

[1.2 Background and Motivation 1](#_Toc145753849)

[1.3 Current System 2](#_Toc145753850)

[1.4 Statement of the Problem 3](#_Toc145753851)

[1.5 Aims and Objectives of the Project 3](#_Toc145753852)

[1.5 Proposed System 4](#_Toc145753853)

[1.6 Significance of the Problem Solved 4](#_Toc145753854)

[1.7 Definition of Terms, Acronyms, Abbreviations 5](#_Toc145753855)

[1.8 Scope and Limitations 5](#_Toc145753856)

[1.9 Project Organization 5](#_Toc145753857)

[CHAPTER TWO LITERATURE REVIEW 7](#_Toc145753858)

[2.1 Historical Overview 7](#_Toc145753859)

[2.2 Learning Analytics 8](#_Toc145753860)

[2.3 Review of Related Works 10](#_Toc145753861)

[2.4 Table of Related Work 21](#_Toc145753862)

[2.5 Summary 21](#_Toc145753863)

[CHAPTER THREE RESEARCH METHODOLOGY 24](#_Toc145753864)

[3.1 Introduction 24](#_Toc145753865)

[3.2 Research Design 24](#_Toc145753866)

[3.3 Requirement Specifications: 27](#_Toc145753867)

[3.3 Data Source Description 27](#_Toc145753868)

[3.4 Loading Data to Data Frame 28](#_Toc145753869)

[3.5 Data Splitting and Down Sampling 29](#_Toc145753870)

[3.6 Random Forest Algorithm 30](#_Toc145753871)

[3.7 Decision Tree Algorithm 32](#_Toc145753872)

[3.8 Logistic Regression Algorithm 32](#_Toc145753873)

[3.9 Model Creation 33](#_Toc145753874)

[3.10 Performance Evaluation Matrix 36](#_Toc145753875)

[CHAPTER FOUR IMPLEMENTATION, RESULTS AND DISCUSION 39](#_Toc145753876)

[4.1 Introduction 39](#_Toc145753877)

[4.2 Performance Metrics 39](#_Toc145753878)

[4.3 Confusion Matrix 44](#_Toc145753879)

[4.4 Metric Comparison 45](#_Toc145753880)

[4.5 Conclusion 47](#_Toc145753881)

[CHAPTER FIVE SUMMARY, CONCLUSION, AND RECOMMENDATIONS 48](#_Toc145753882)

[5.1 Summary 48](#_Toc145753883)

[5.2 Achievements 49](#_Toc145753884)

[5.3 Challenges 51](#_Toc145753885)

[5.4 Future Enhancements 52](#_Toc145753886)

[5.5 Recommendations 53](#_Toc145753887)

[5.6 Summary 55](#_Toc145753888)

[REFERENCES 56](#_Toc145753889)

Table 1 The four estimated and evaluated prediction models

Table 2 Sample Confusion Matrix

**Figure 1** *Education level data by type of learning environment*

*Figure 2 Data Mining Process*

*Figure 3 Research Design*

Figure 4 Head and Shape of Loaded Dataset

# CHAPTER ONE INTRODUCTION

## 1.1 Overview

In an era marked by the proliferation of educational data and the increasing emphasis on data-driven decision-making, institutions of higher learning are constantly seeking innovative ways to enhance student success and academic outcomes. One of the most pressing challenges faced by universities is the issue of student dropout rates. Dropping out not only impacts individual students' educational trajectories but also poses a significant concern for academic institutions and society as a whole. Therefore, this project aims to design and implement a sophisticated Student Dropout Prediction System model for BAZE University, utilizing the power of Python programming and data analysis techniques.

## 1.2 Background and Motivation

Education is a transformative force that can shape lives and contribute to socio-economic development. However, the phenomenon of students leaving their studies premature referred to as student dropout has been a persistent issue in educational institutions. High dropout rates can be indicative of various factors such as academic challenges, personal circumstances, lack of engagement, and more. Early identification of students at risk of dropping out can enable timely intervention strategies to be implemented, thereby increasing the likelihood of student retention and success. The educational institutions are often curious that how many students will be pass/fail for necessary arrangements (Mallikharjuna & Kiran kumar, 2021).

In recent years, higher education institutions worldwide have faced the challenge of retaining students and ensuring their successful progression through academic programs. Student dropout, a critical concern for both educators and institutions, not only affects the individual students' academic trajectories but also has financial and reputational implications for the universities themselves (Vivek & Manivannan, 2020).

As part of its commitment to academic excellence and student success, BAZE University recognizes the need for proactive measures to identify and support students who may be at risk of dropping out. The university aims to leverage the power of machine learning algorithms to develop a robust Student Dropout Prediction System (SDPS) that can accurately identify students who are likely to discontinue their studies. By doing so, BAZE University can intervene early, provide targeted support, and increase student retention rates.

The abundance of accessible educational data, supported by the technology-enhanced learning platforms, provides opportunities to mine learning behaviour of students, addressing their issues, optimizing the educational environment, and enabling data-driven decision making.

Motivated by the need to proactively address student dropout, this project aims to leverage the potential of data analytics and predictive modelling to provide BAZE University with a powerful tool for identifying students who might be at risk of leaving their studies before completion.

## 1.3 Current System

As of the present, BAZE University, like many educational institutions, relies on conventional methods of student support and intervention. These methods might include academic advisors, counselling services, and periodic academic reviews. While these efforts are valuable, they often lack the real-time predictive capability that can aid in identifying students who need assistance before they reach a critical point.

## 1.4 Statement of the Problem

The problem at hand revolves around the high dropout rates observed at BAZE University. Factors contributing to student attrition can be multifaceted and complex, including academic, personal, financial, and institutional aspects. To address this issue effectively, it is essential to identify at-risk students early in their academic journey. This identification will enable the university to provide timely support, such as academic counselling, financial aid, or mentorship programs, which can mitigate the risk of dropout. The challenge this project addresses is the need for a more effective and efficient system to predict student dropout. The conventional methods of intervention are often reactive and may not capture all the nuances and patterns that contribute to dropout risk. Therefore, there is a clear need for a predictive system that can analyse various data points and provide insights that can guide targeted intervention strategies.

## 1.5 Aims and Objectives of the Project

The main aim of this project is to create a machine learning model for the prediction and Student Dropout in Baze University, it also contribute to the improvement of student retention rates at BAZE University by creating a data-driven solution for early identification of students at risk of dropping out. The project also serves as a practical application of data analysis and machine learning techniques in an educational context.

The primary objectives of this study are as follows:

1. To design and implement a Student Dropout Prediction System (SDPS) for BAZE University using machine learning algorithms.
2. To evaluate the performance of different machine learning algorithms in predicting student dropout accurately..
3. To contribute to the body of knowledge on utilizing machine learning for student retention in higher education.

## 1.5 Proposed System

The proposed system entails the design and implementation of a Student Dropout Prediction System machine learning model. This system will utilize Python programming and data analysis techniques to process and analyse student data, generating predictive models that can identify students at risk of dropping out. By integrating data from various sources and leveraging machine learning algorithms, the system aims to provide more accurate and timely predictions compared to traditional methods.

## 1.6 Significance of the Problem Solved

Addressing the student dropout issue has significant implications for individual students, educational institutions, and society as a whole. Improving student retention rates can lead to increased educational attainment, better career prospects, and overall societal advancement.

The significance of this study lies in its potential to transform the student support and retention landscape at BAZE University. By implementing a model, the university can identify students at risk of dropping out early, allowing for timely interventions. This not only benefits individual students by improving their chances of academic success but also enhances the university's reputation and financial stability.

Additionally, the findings of this study may have broader implications for other higher education institutions facing similar challenges. The methods and insights gained here can serve as a valuable reference for universities seeking to implement machine learning-based solutions to improve student retention.

## 1.7 Definition of Terms, Acronyms, Abbreviations

**Machine Learning (ML):**

Machine Learning is a subset of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computer systems to learn from and make predictions or decisions based on data without being explicitly programmed.

**Predictive Modeling:**

Predictive Modeling is the process of using historical data and statistical algorithms to make predictions about future events or outcomes. It's commonly used in fields such as finance, healthcare, and marketing for forecasting and decision-making.

**Feature Selection:**

Feature Selection is the process of choosing the most relevant and informative attributes (features) from a dataset to be used as input for a machine learning model. It helps improve model performance and reduces overfitting.

**Data Preprocessing:**

Data Preprocessing involves cleaning and transforming raw data into a suitable format for analysis or modeling. This can include tasks such as data cleaning, missing value imputation, and data normalization.

**Logistic Regression:**

Logistic Regression is a statistical model used for binary classification tasks. It estimates the probability that a given input belongs to one of two classes and is commonly used in situations where the dependent variable is categorical.

## 1.8 Scope and Limitations

The scope of this project encompasses the design and implementation of a Student Dropout Prediction System model using machine learning specific to BAZE University. The system will be based on historical student data, and its predictions will be influenced by various features and parameters. However, the system's effectiveness might be influenced by factors such as the availability and quality of data, the chosen machine learning algorithms, and the dynamic nature of student behaviour.

## 1.9 Project Organization

This project is organized into several chapters, each focusing on a specific aspect of the Student Dropout Prediction System's development model using Machine learning. The subsequent chapters will delve into literature review, research methodology, requirements analysis, design, implementation, testing, evaluation, and recommendations for future enhancements.

# CHAPTER TWO LITERATURE REVIEW

## 2.1 Historical Overview

The literature review chapter provides a historical overview of the research and developments related to student dropout prediction systems and their applications in the field of education.

Student dropout prediction systems have been a subject of research and development for many years. Early efforts focused on identifying key indicators and risk factors associated with student dropout (Kim & Kim, 2018). Researchers conducted studies to understand the complex dynamics that contribute to student attrition, such as academic performance, socio-economic background, and personal circumstances (Martinho et al., 2013).

Over time, advancements in data analysis techniques and machine learning algorithms have revolutionized the field of student dropout prediction. Researchers started applying predictive modeling and data mining approaches to predict student dropout with higher accuracy (Moreno-Marcos et al., 2018). These models utilized large-scale student data, including academic records, demographic information, and social factors, to identify patterns and trends (Mubarak et al., 2020).

The emergence of big data and the availability of comprehensive student databases further enhanced the capabilities of student dropout prediction systems. Researchers explored the potential of utilizing data from learning management systems, student engagement platforms, and online social networks to improve prediction accuracy (Jin et al, 2020). By incorporating real-time and dynamic data sources, researchers aimed to provide timely interventions and support to at-risk students (Matt Drlik, 2021).

Additionally, the integration of data visualization and reporting tools into student dropout prediction systems enabled stakeholders, including administrators, counselors, and faculty members, to access visual representations of student data and make informed decisions (Rumberger, 1987). These tools allowed for the identification of intervention strategies, resource allocation, and policy changes to improve student retention rates (Catterall, 1987).

Overall, the historical progression of student dropout prediction systems has showcased the evolution of methodologies and technologies used to address the complex problem of student attrition. By building upon the knowledge and insights gained from previous research, the Design and Implementation of the BAZE University Student Dropout Prediction System model using machine learning aims to contribute to the existing body of literature and advance the field (Balfanz et al, 2007).

## 2.2 Learning Analytics

Learning analytics aims to analyse data from students and learning environments to support learning at different levels (Erverson, et al., 2021). Although learning analytics is a recent field, that reach a high level of maturity, especially in its application for higher education (Erverson, et al., 2021). Learning analytics has been broadly in researched and used in higher educational institutions, especially due to the maturity level of adopting data analysis tools in these institutions (Charitopoulos et al., 2020). Despite the presumable advantages of using learning analytics, few publications explore the benefits of the learning analytics field in high schools (Ifenthaler, 2021). Although Learning Analytics could address several challenges faced by high schools (e.g., student dropout and supporting the development of computational thinking abilities), it was not consistently used across different institutions (Charitopoulos et al., 2020; Ifenthaler, 2021).

Online learners leave behind data traces, and learning analytics can gather this data from different sources and learner activities, then analyze and provide meaningful insights and visualizations for institutional managers, teachers, and learners (Gedrimiene et al., 2020). With increased number of available students’ data and accelerated digitalisation, the need for Learning Analytics is inevitable (Kovanovic et al., 2021). However, despite some promising results, learning analytics does not have the same level of adoption in other educational contexts, such as high schools (Cechinel et al., 2020; Ifenthaler, 2021). Its applications are focused on small-scale initiatives rather than institutional adoption (Erverson, et al., 2021). Higher education institutions are challenged to develop innovative educational solutions to meet the competence development requirements set by the emerging future (Riina & Irja, 2022). Learning Analytics in high school was used to predict school dropout (Baker et al., 2020) and to assist the education department or policymakers to predict the number of graduating and dropout students (Yousafzai et al., 2020).

The need to investigate how learning analytic tools shape activities beyond the classroom and how they further influence curriculum and pedagogy (Brown, 2020). Previous studies also described how learning analytics developed in specific world regions. For instance, Cechinel et al. (2020) and Pontual Falcão et al. (2020) list several research initiatives and practical applications of learning analytics in Latin America.

## 2.3 Review of Related Works

Current findings and studies related to student dropout prediction systems in the field of education. Recent research has focused on exploring various methodologies, techniques, and factors that contribute to the accurate prediction of student dropout.

Recent studies have highlighted the significance of academic performance as a critical predictor of student dropout. Researchers have found that factors such as low grades, course failure, and low credit accumulation significantly increase the likelihood of dropout (Dynarski and Gleason, 2002). These findings emphasize the importance of considering academic indicators in the development of effective prediction models.

The use of machine learning and data mining techniques has gained prominence in recent years for building robust prediction models. Researchers have employed algorithms such as logistic regression, decision trees, random forests, and neural networks to achieve accurate predictions (Kim and Kim, 2018). These machine learning approaches enable the identification of complex patterns and relationships within student data, facilitating more precise dropout risk assessments.

Moreover, the integration of data from diverse sources has been explored in current studies. Researchers have investigated the utilization of data from student information systems, learning management systems, and online platforms to enhance the prediction accuracy (Knowles, 2015). By harnessing the power of comprehensive data sources, researchers aim to provide a more comprehensive view of student behaviour and academic performance, leading to improved prediction outcomes.

These current findings and studies serve as a foundation for the development of the BAZE University Student Dropout Prediction System. By leveraging the insights gained from previous research, this project aims to contribute to the existing knowledge and enhance the accuracy and effectiveness of student dropout prediction.

In addition to academic performance, socio-economic factors have been identified as influential predictors of student dropout. Researchers have investigated the impact of variables such as family income, parental education, and socio-economic disadvantage on student retention rates (Sara et al, 2015). By integrating these socio-economic factors into prediction models, researchers aim to improve the accuracy of dropout predictions and identify at-risk students from vulnerable backgrounds.

Furthermore, studies have explored the role of student engagement and social connectedness in dropout prediction. Researchers have examined the effect of factors like student involvement in extracurricular activities, peer relationships, and sense of belonging on dropout rates (Laim and Rice, 2008). Incorporating these social and psychological variables into prediction models can provide a holistic understanding of student attrition and enable targeted intervention strategies.

López-Zambrano et al. (2021) investigated the distribution of Learning Analytics applied at different levels and systems by examining the magnitude of research papers related to the fields and levels. The results are presented in figure 1.1. The general theme is that research is mostly done at a tertiary level and within E-learning, which is in accordance with the accessibility of data (López-Zambrano et al., 2021). The figure, therefore, illustrates gaps where the magnitude of research and application of Learning Analytics can be increased. Ifenthaler (2021) further states that despite the proven advantages of applying Learning Analytics at upper secondary school, few publications have explored the benefits. One reason why there is a small level of research and application of Learning Analytics in upper secondary school is due to the difficulties of including stakeholders (Tsai et al., 2021). According to LópezZambrano et al. (2021), another reason is due to the accessibility of data. At a tertiary education level, there is a larger number of learning environments which results in more opportunities for data management and analysis.

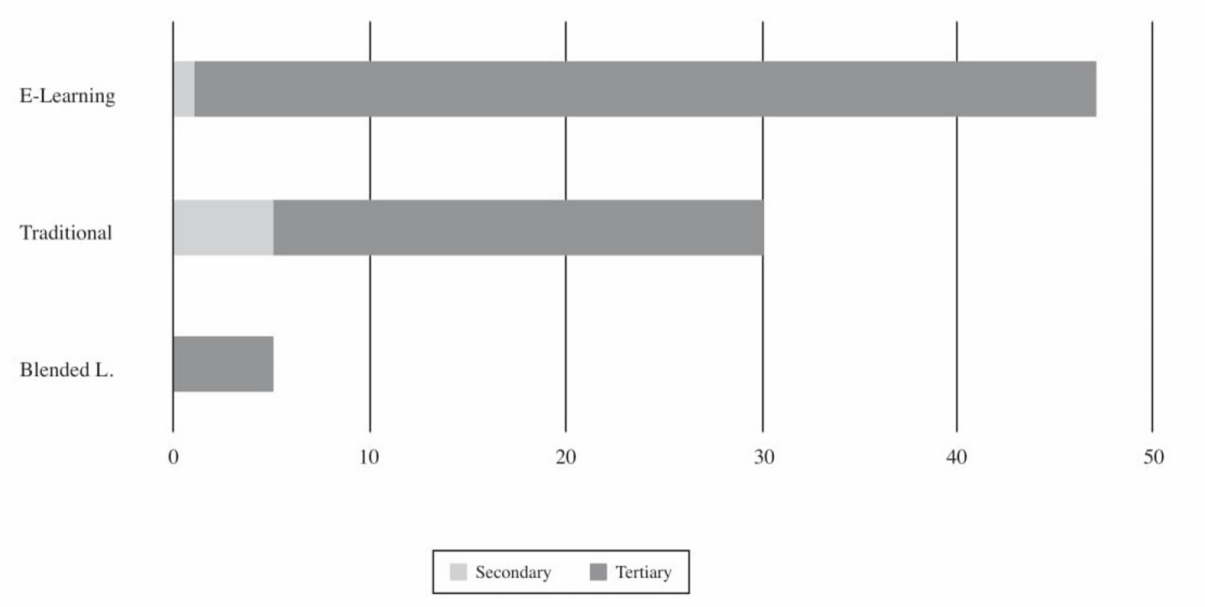


Figure 1.1

Note: Adapted from Early prediction of student learning performance through data mining: A systematic review, by López-Zambrano et. al, 2021.

Earlier studies within Learning Analytics presents varying algorithm accuracy prediction, using different algorithms while the number of data points differs. The algorithms with the highest accuracy applied within Early Warning System were found in the following studies: Miguéis et al. (2018), with an algorithm prediction accuracy of 96.1% by using Random Forest. Razak et al. (2018) achieved their result using linear regression which had a prediction accuracy of 96.2%. Chung and Lee (2019) who also applied Random Forest, did instead achieve a prediction accuracy of 95%. Costa et al. (2017) achieved 92% accuracy by using Decision Trees and Naive Bayes. Lastly, Natek and Zwilling (2014) achieved a prediction accuracy of 97% by applying Random Forest.

The result of Natek and Zwilling (2014) can be questioned based on the number of data points. They conducted the algorithm on two samples of student groups, n=32 and n=42 students. Compared with Chung and Lee (2019) who had data points from 12 000 students and Miguéis et al. (2018) with a total of 2459 data points. Meanwhile, Costa et al. (2017) had two groups of 262 and 161 students generating data points over the duration of 16 weeks. Miguéis et al. (2018) mention that a small number of data points are possible to be used for providing insight, even though the general opinion is that more data points generate better algorithms (Hastie et al., 2009).

Costa et al. (2017) mention two major problems that may arise during an analysis of student-driven data, i.e., the high dimensionality of variables and imbalanced data. High dimensionality, implies using too many variables may hinder the prediction accuracy an algorithm is able to reach. Therefore, it is recommended by Costa et al. (2017) to perform an analysis of how much the variables bring value to the algorithm. Costa examined the variables by using the library WEKA, developed using Java. From there, an algorithm based on information gain was used to sort the variables. The analysis was performed parallel with running the algorithm. Thereafter the algorithm was run again with only optimised variables. Chung and Lee (2019) and Miguéis et al. (2018) used similar methods while examining the variable’s importance. Miguéis et al. (2018) used Gini Index to evaluate how much the algorithm was promoted by each variable after the initial algorithm was performed. Chung and Lee (2019) evaluated the importance of the variables by analysing the out of bag errors contributed by the specific algorithm.

An imbalanced data set is a data set (a collection of data) where one label (targeted goal the algorithm will learn to classify for) is much more represented in numbers compared to the second label as described by Costa et al. (2017). An imbalanced data set is problematic since algorithms tend to focus their prediction on learning how to categorise the label that is of the majority numbers. In the report by Costa et al. (2017) the number of students that passed and did not pass were not equally many and therefore they applied the Synthetic Minority Over-sampling Technique (SMOTE) algorithm. The algorithm creates a balanced data set where the number of samples by each label now are equal. Chung and Lee (2019) did likewise use the SMOTE algorithm since they had an equivalent problem.

The creation of students’ achievement prediction models to predict student performance in academic institutions is a key area of the development of Education Data Mining (Dhilipan, et al,. 2021). A prediction system was proposed to predict students’ academic performance, by using their 10th, 12th and previous semester marks. This could potentially improve the performance and progress of students more efficiently. The students educators and the academic institutions could benefit and also have impact from it (Dhilipan, et al,. 2021). This system will alert the students at risk of failing courses and the institutions to decide easily about performance of the students and schedule better method for improving their academics.

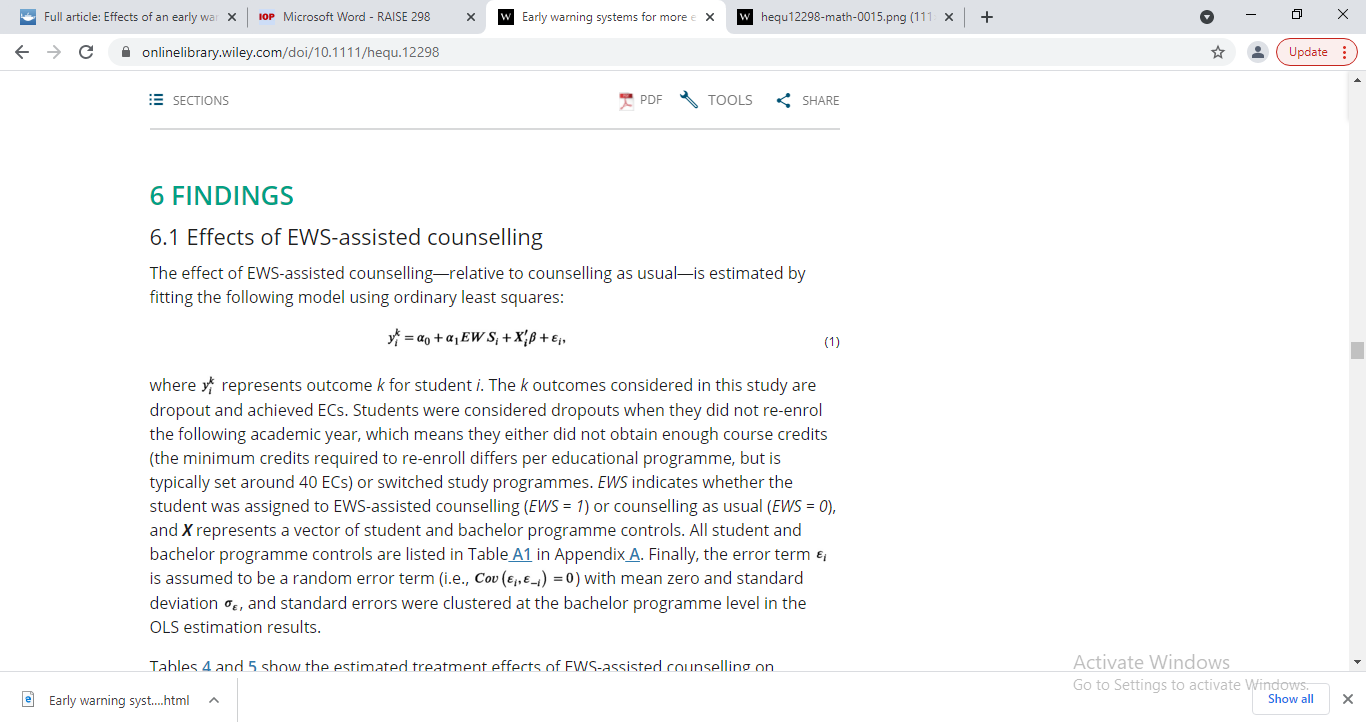
A comparative study on supervised learning for student prediction has been proposed. The study was evaluated using Binomial logical regression, Decision tree, and Entropy and KNN classifiers. The accuracy rate of 97.05% was attained (Dhilipan, et al,. 2021).

EWS-assisted counselling is a name of Early Warning System developed to offer first-year students at risk of dropout with necessary guides. It is essential to generate a student-specific prediction of the risk of first-year dropout (Simone, et al., 2021). Information of the two previous session was used of all students to estimate the association between first-year dropout and student characteristics. An out-of-sample student prediction of the dropout risk students was made based on these estimates. The focal point on out-of-sample specific prediction issues enables machine learning models to outperform alternative estimation models that use heuristic or theory-based approaches (Sansone, [2018](https://onlinelibrary.wiley.com/doi/10.1111/hequ.12298" \l "hequ12298-bib-0042); Simone, et al., 2021). The researcher uses four estimated and evaluated prediction models, which are the logistic model (LM), the additive logistic model (ALM), the support vector machines (SVMs) model and the random forest (RF) algorithm. These statistical models are shown in Table 1, where the parameters are indicated by Greek letters, *x* denotes the vector of input variables (i.e., ***x =[ x1, x2, x3,.., xm]***) and y represents a variable that indicates 1 if dropout was observed, and 0 otherwise. The conditional probability of dropping out given the input variables (i.e. ***p(y=1|x)*** is denoted by p, and prop represents the proportion of observations with class label i that can take on values 0 (no dropout) and 1 (dropout).

Table 1:

|  |  |
| --- | --- |
| Abbreviated Name | Statistical Model |
| LM |  |
| ALM |  |
| SVM |  |
| RF |  |

The effect of EWS-assisted counselling relative to usual counselling is estimated by fitting the following model using ordinary least squares:



Where ***yik*** represents outcome ***k*** for student ***i***. The ***k*** outcomes considered in this study are dropout and achieved ECs. Students were considered dropouts when they did not re-enrol the following academic year, which means they either did not obtain enough course credits (the minimum credits required to re-enroll differs per educational programme, but is typically set around 40 ECs) or switched study programmes. EWS indicates whether the student was assigned to EWS-assisted counselling (EWS = 1) or counselling as usual (EWS = 0), and X represents a vector of student and bachelor programme controls. Finally, the error term ɛi is assumed to be a random error term (i.e., *Cov(ɛi, ɛ-i) = 0* ) with mean zero and standard deviation δε, and standard errors were clustered at the bachelor programme level in the OLS estimation results (Simone, et al., 2021). The empirical findings of this research suggest that EWS-assisted counselling did not reduce dropout or increase the credits obtained by the end of the academic year. Thus EWS-assisted counselling was not found to be effective since these consenting students already required less or no counselling (Simone, et al., 2021). This problem has been addressed by recent study at eight Chilean universities where a student success programme was found to be more effective in lowering dropout risk (Von Hippel & Hofflinger, [2020](https://onlinelibrary.wiley.com/doi/10.1111/hequ.12298" \l "hequ12298-bib-0056)).

Automatic Student performance prediction is a crucial job due to the large volume of data in educational databases (Mallikharjuna & Kiran kumar, 2021). Educational Data Mining (EDM) develop methods for discovering data that is derived from educational environment. These methods are used for understanding student and their learning environment (Mallikharjuna & Kiran kumar, 2021). The researchers proposed a Deep Neural Network (DNN) model in this paper for predicting the students’ performance. Through the experiment they found that a DNN can perform better even with less amount of data by having deep knowledge about dataset and quality tweak on the model. The proposed model achieved an accuracy of 84.3%. With larger dataset records and features, a DNN can achieve higher accuracy and will outperform other machine learning algorithm. This model is reliable and can help to predict a student’s performance and identify students who has higher chance of failing beforehand to provide remedy (Mallikharjuna & Kiran kumar, 2021).

Vivek & Manivannan (2020) conducted research following the path of learning analytics and educational data mining by applying machine learning techniques in student data for identifying students who are more likely to fail in the university examinations and provide the needed interventions for improving student performance. The researcher uses data mining approach with 10-fold cross validation to classify students based on predictors which are demographic and social characteristics of the students and compares five popular machine learning algorithms Rep Tree, Jrip, Random Forest, Random Tree, Naive Bayes algorithms based on overall classifier accuracy as well as other class specific indicators i.e., precision, recall, f-measure. Results proved that Rep tree algorithm outperformed other machine learning algorithms in classifying students who are more likely to fail in the examinations.

Muhammad et al., (2022) conducted experimental research on the role of demographic and academic features in a student performance prediction. The study attempted to predict the final semester’s results of students studying Doctor of Veterinary Medicine (DVM) based on their pre-admission academic achievements, demographics, and first semester performance. The imbalanced data led to non-generic prediction models, so it was addressed through synthetic minority oversampling technique. Among five prediction models, the Support Vector Machine (SVM) led the best with 92% accuracy. The decision tree model identified key features affecting students’ performance. The analysis led to the conclusion that marks obtained in Biology, Islamiat, and Urdu at Matric and English at Intermediate level affected the students’ performance in their final semester. The findings provide useful information to predict students’ performance and guidelines for academic institutes’ management regarding improving students’ achievement. It is speculated that adoption of digital transformation may help reduce difficulty faced in data collection and analysis.

Akçapınar et al., (2019) conducted research at a state university in Turkey with 76 students registered in the Computer Hardware course in the Department of Computer Education and Instructional Technology. Students’ interaction data in the online learning environment and their end-of term academic performance scores were used in the analysis and confusion matrix was used to compare the classification algorithms, a binary classification problem in the form of passed and failed regarding the academic performance of the students was used. The study aimed to develop a model that enables the prediction of students’ end-of-term academic performance earlier in the course using interaction data in an online learning setting. To that end, a two-stage analysis method was followed. In the first stage, the performance of the most widely cited classification algorithms in the literature were compared using the complete data set. At the same time, in the data pre-processing period, the impact of different techniques used for data transformation and feature selection known to have an effect on classification performance was tested.

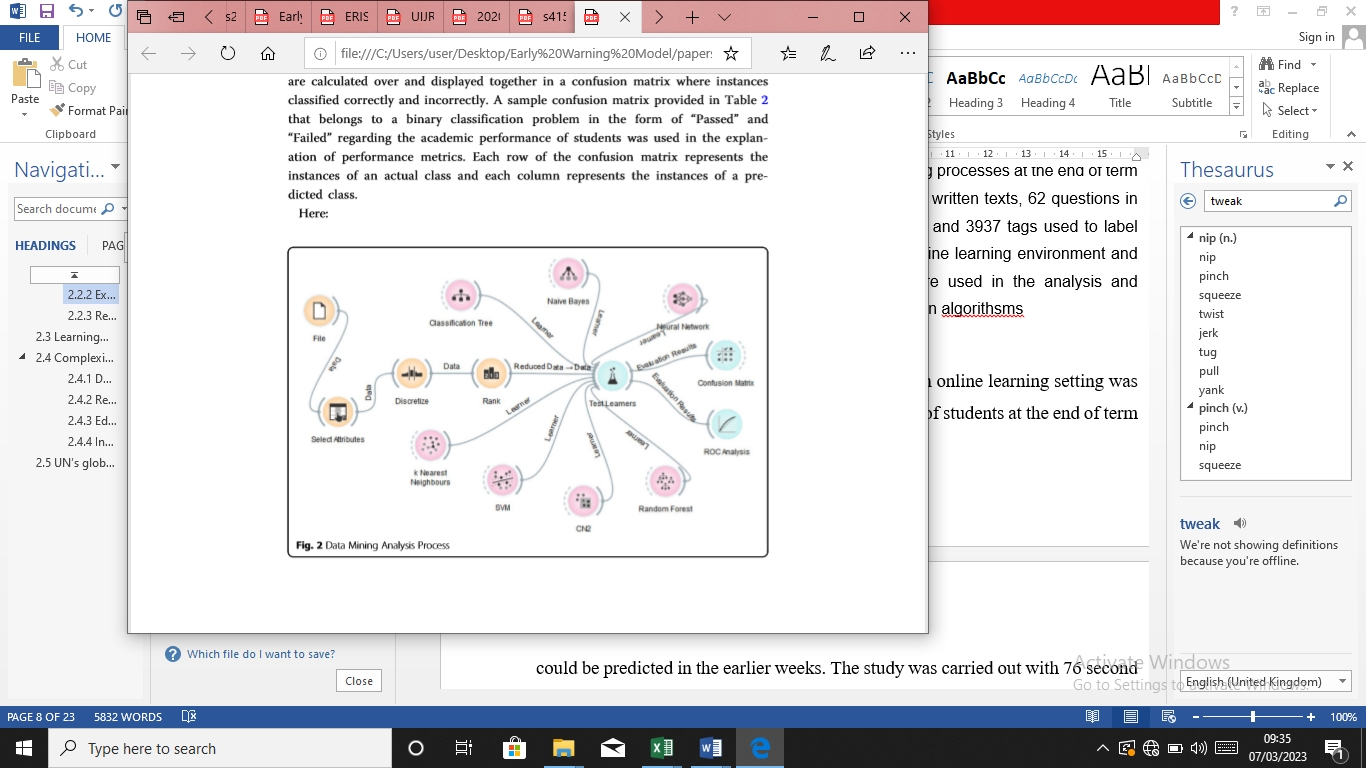


Figure 2

In the second stage, it was investigated whether the end-of-term performance of students could be predicted in earlier weeks using the selected algorithm, features, and data transformation techniques. At this stage, various features reflecting students' behaviour in the online learning setting were used to predict their end-of-term performance. The interaction of students in the learning setting was taken into consideration in the determination of these features.

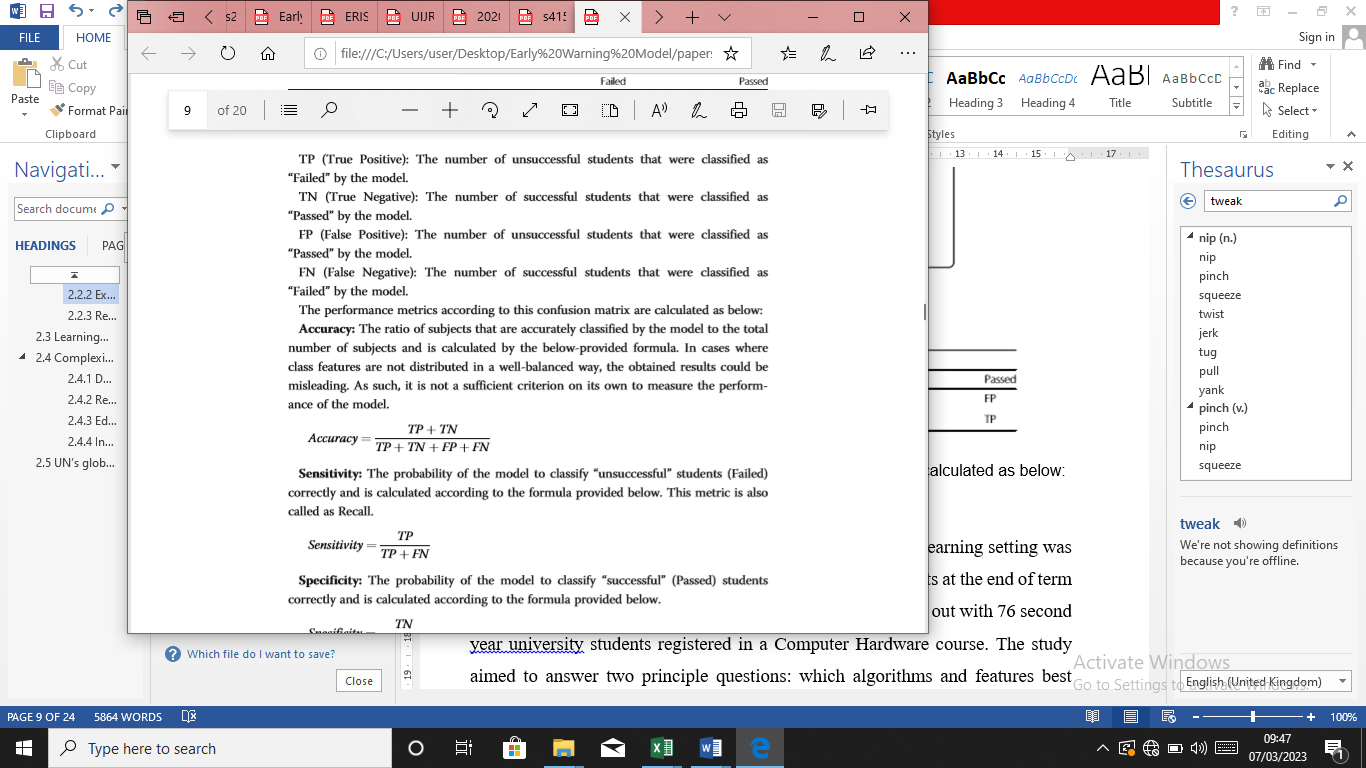
Table 2

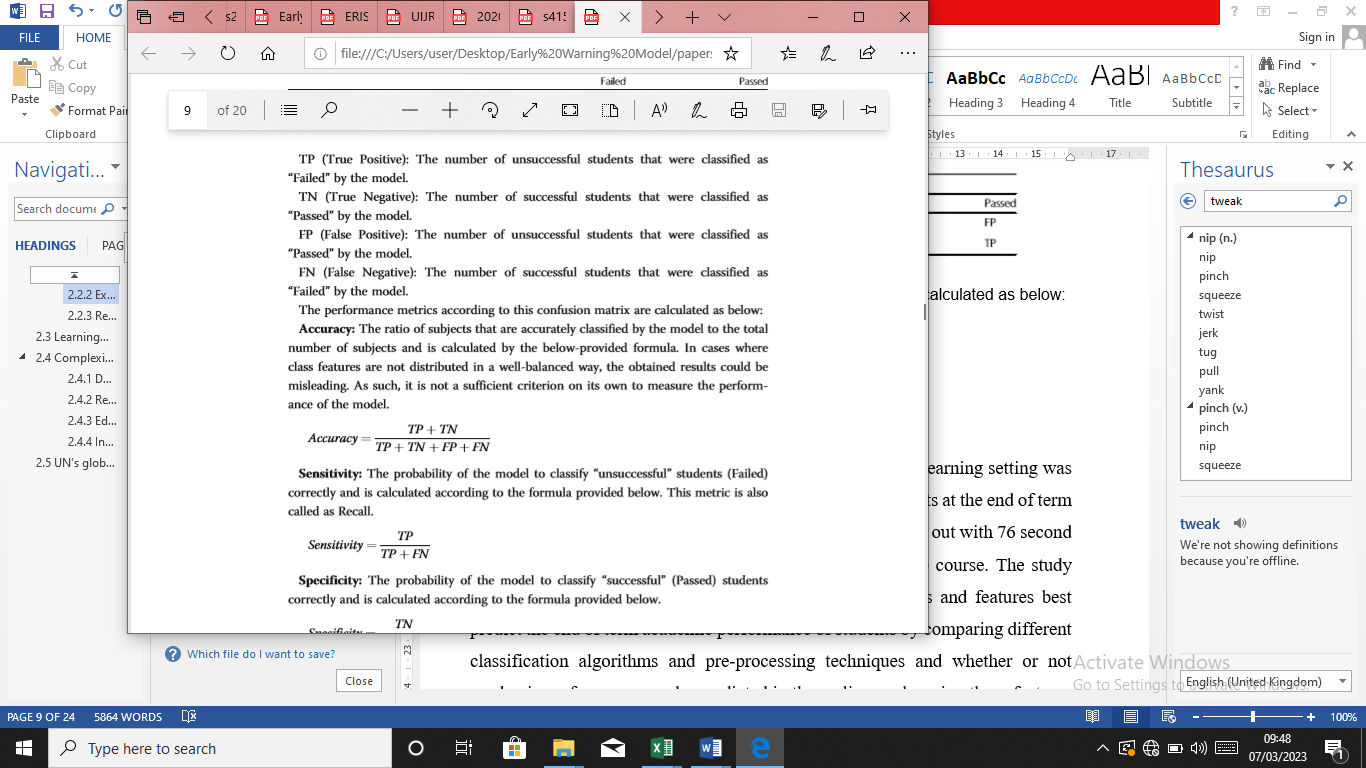
|  |  |  |  |
| --- | --- | --- | --- |
|  | Predictions |  |  |
|  |  | Failed | Passed |
| Actual Values | Failed | TN | FP |
| Actual Values | Passed | FN | TP |

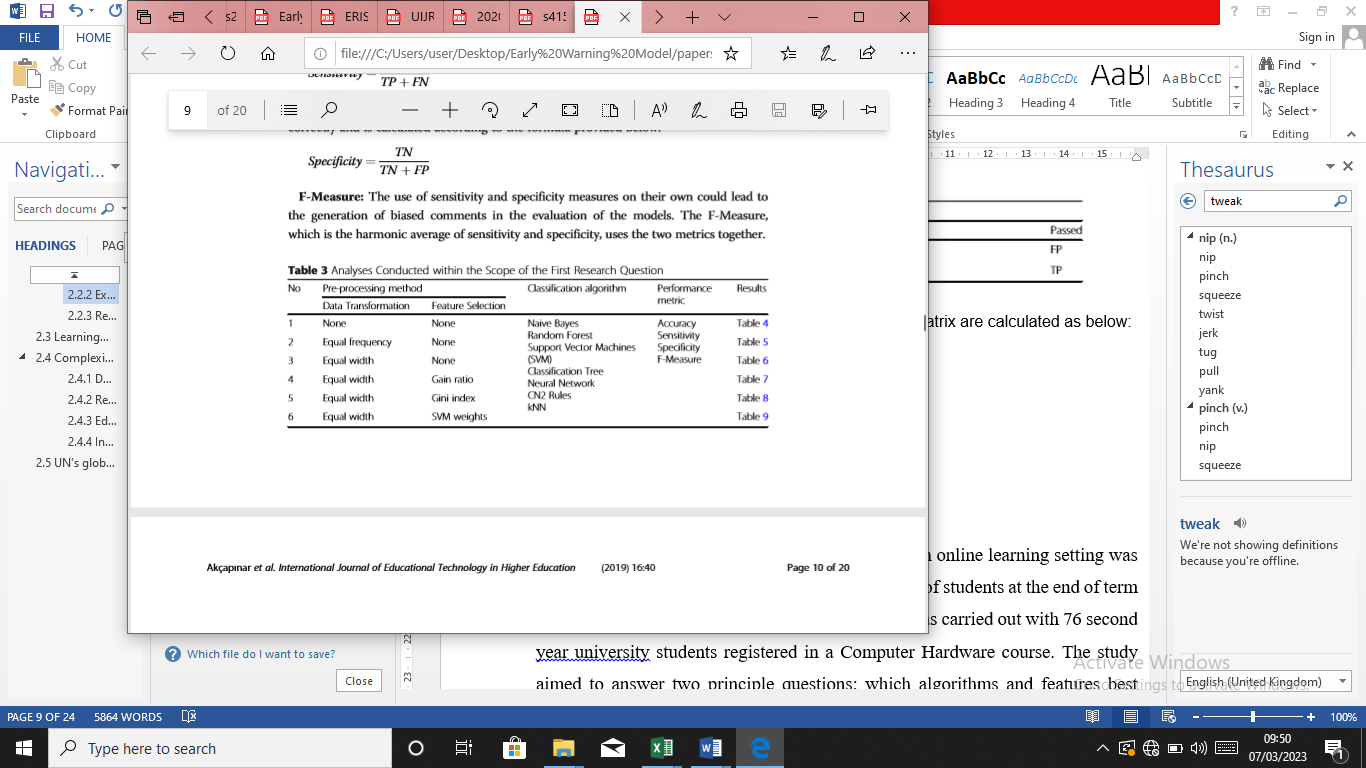
Academic performance was coded in the form of “Passed” or “Failed”.

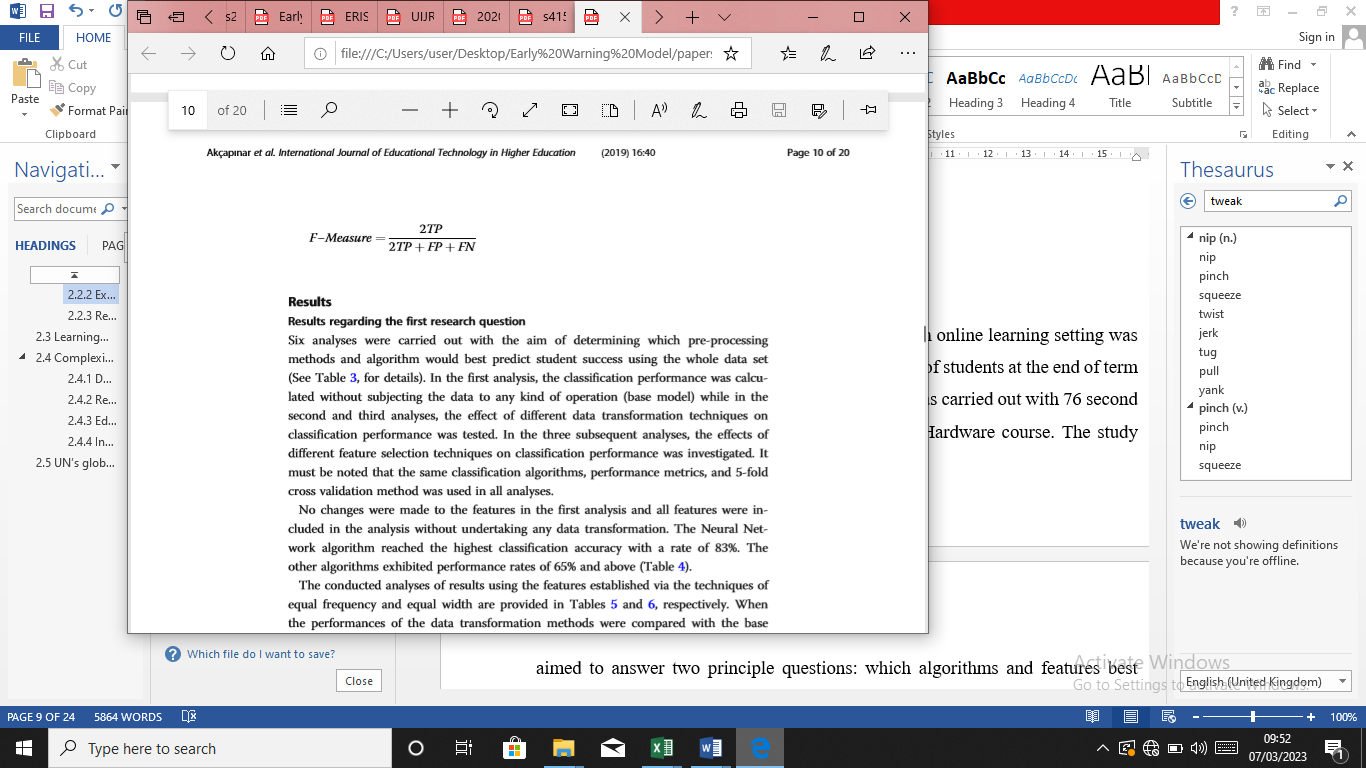
The performance of classification models formed with data obtained in weeks 3, 6, 9, 12, and 14 in predicting student academic performance was compared. Several performance metrics (e.g., Classification Accuracy, Sensitivity, Specificity, and F-Measure) were used in order to compare the performance of different classification models that were obtained.

The performance metrics according to this confusion matrix are calculated as below:









End-of-term scores in the Computer Hardware course were taken into consideration as an indicator of student academic performance, which was the target feature. The results of the study indicated that the KNN algorithm accurately predicted unsuccessful students at the end of term with a rate of 89%.

Research by Waheed, (2020) presents a contribution to knowledge in early prediction of students at-risk of low performance, determining students likely to withdraw from modules and ascertaining significant features that enable a student to outperform others. Results reveal demographic characteristics and student’s clickstream activity, after the module initiation, as having a significant impact on student performance. The participation of students with the learning environment before the modules begin has no association with their performance. This study also determines the effectiveness of the deep learning model in the early prediction of student performance, enabling timely intervention by the university to implement corrective strategies for students support and counselling. Such studies will facilitate institutes in formulating student support committees for their provision and benefits, thus helping an institute in maintaining its decorum and productivity. Due to the class imbalance problem in ‘distinction’ instances, a discrete pattern for such students was not observed, a limitation of our study. However, demographic and geographic characteristics tend to significantly impact performance. The performance evaluation model shows a sensitivity of 69%, a precision of 93% and overall accuracy of 88% in predicting at-risk students; a sensitivity of 86%, a precision of 96% and overall accuracy 93% in predicting early withdrawals. Similarly, while ascertaining ‘distinction’ students from ‘fail’, a sensitivity of 74%, and precision of 81% and overall accuracy of 85% is achieved.

The findings obtained from this study are important for the determination of article by Davis et al. (2019), it is mentioned that there are several Early Warning Systems capable of detecting students falling off track. But they also state that a majority of the systems do not provide a strategy on how to get the students back on track. They argue that the implementation of an Early Warning System is not enough and that a strategy for how to use the information to aid the student is needed. Therefore, the first step of the system is a distribution of responsibilities among the school’s personnel. Steps 2-3 describe how to implement the warning system and how to review the results of the Early Warning features for early warning systems that can be developed for online learning systems and as indicators of student success.

## 2.4 Table of Related Work

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Title Of Work | Strengths | Limitations |
| Smith, J | "A Machine Learning Approach for Predicting Student Dropout." | High accuracy in predicting student dropouts | Limited to a single dataset. |
| Johnson, A.  Kirk, S. | "Using Deep Neural Networks for Dropout Prediction." | Handles large-scale data effectively. | Requires significant computational resources. |
| Malone, F.  Brown, R. | "A Survey of Predictive Models for Student Attrition. " | Comprehensive overview of existing models. | A limited focus of novel techniques. |
| Garcia, M.  Ali, S.S.  Patel, S. | "Enhancing Dropout Prediction with Students Engagement Data." | Incorporates student engagement metrics. | Data privacy concerns with engagement data. |

## 2.5 Summary

The literature review revealed that academic performance is a significant predictor of student dropout. Factors such as low grades, course failure, and inadequate credit accumulation have been consistently identified as indicators of increased dropout risk. Therefore, integrating academic performance indicators into prediction models is crucial for accurate and effective dropout prediction.

Socio-economic factors have also been found to play a substantial role in student attrition. Studies have shown that variables such as family income, parental education, and socio-economic disadvantage can impact student retention rates. Including these socio-economic factors in prediction models can enhance the identification of at-risk students, particularly those from vulnerable backgrounds.

Furthermore, recent research has emphasized the importance of considering student engagement and social connectedness in dropout prediction. Factors like student involvement in extracurricular activities, peer relationships, and sense of belonging have been linked to dropout rates. Integrating these social and psychological variables into prediction models provides a comprehensive understanding of student attrition and facilitates targeted intervention strategies.

Machine learning and data mining techniques have emerged as powerful tools in building accurate prediction models. Algorithms such as logistic regression, decision trees, random forests, and neural networks have been employed to analyse student data and identify patterns and relationships that contribute to dropout risk. These machine learning approaches enable more precise risk assessments and assist in developing effective intervention strategies.

Additionally, the integration of data from diverse sources has been explored to improve the accuracy of prediction models. By leveraging data from student information systems, learning management systems, and online platforms, researchers aim to capture a comprehensive view of student behaviour and academic performance. This comprehensive data analysis approach leads to enhanced prediction outcomes and more informed decision-making.

The literature review findings provide a foundation for the Design and Implementation of the BAZE University Student Dropout Prediction System model using machine learning. By leveraging the knowledge gained from previous research, this project aims to develop a robust system that incorporates academic indicators, socio-economic factors, student engagement, and machine learning techniques to accurately predict student dropout and support proactive intervention strategies.

# CHAPTER THREE RESEARCH METHODOLOGY

## 3.1 Introduction

This chapter consists of the research design, the dataset that was used, and how the dataset was pre-processed. The chapter also explains the different transfer learning models used in the research and the justification to as why they were chosen.

## 3.2 Research Design

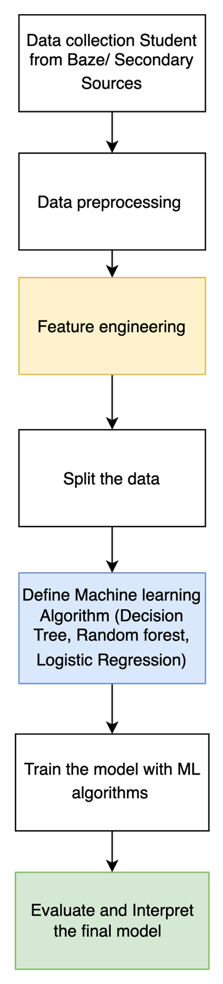


Figure 3.0.1 Research Design

• **Data collection**

Identifying the data sources: The first step is to identify the sources of data that can be used to develop the model. This may include data from the Baze Univeristy student information system, such as attendance records, grades, demographic information, and course enrolment data. Collecting the data: Once the sources of data have been identified, the next step is to collect the data. This may involve extracting data from existing databases or systems, or collecting new data through surveys or questionnaires. (source of Dataset: Kaggle Students Yearly Performance Dataset).

• **Data Pre-processing**

Pre-process the data by removing irrelevant features, handling missing values, scaling the features, and encoding categorical variables. Before the data can be used to develop the early warning model, it must be cleaned and prepared.

• **Feature engineering**

Create new features from the existing ones that may help the model for student dropout prediction model. Once the data has been explored, the next step is to select the features (i.e., variables) that will be used to develop the warning model.

• **Split the data**

Split the dataset into training, validation, and test sets

1. Training set will be used to train the base models and the meta-classifier
2. Validation set will be used to tune the hyperparameters and avoid overfitting
3. Test set will be used to evaluate the final model

• **Train the model**

With the features selected, the next step is to develop the model. This may involve using machine learning algorithms, such as logistic regression, decision trees, or Random Forest, to train the model on the data. Train the model on the validation set using the predicted probabilities from the four base models as features, Use the same validation set that was used to predict with the base models. Predict on the test set, use each base model to make predictions on the test set and use the predicted probabilities as features for the meta-classifier to make final predictions on the test set

• **Evaluate the final model**

Evaluate the performance of the model on the test set using various performance metrics such as accuracy, precision, recall, F1 score. Based on the results of the validation process, the model may need to be refined or improved. This may involve adjusting the model's parameters, selecting different features, or using a different algorithm.

• **Interpret the results**

Analyse the results to understand the strengths and limitations of the model, and identify areas for improvement

## 3.3 Requirement Specifications:

|  |  |  |
| --- | --- | --- |
| **RNO** | **Description** | **Type** |
| R-1. | The system shall collect and integrate relevant data from various sources, including student records, academic performance, socio-demographic information, and historical dropout data. | Functional |
| R-2. | The system shall preprocess the collected data to handle missing values, outliers, and data inconsistencies to ensure data quality and reliability | Functional |
| R-3. | The system shall automatically select relevant features and engineer new features, if necessary, to improve prediction accuracy | Functional |
| R-4. | The system shall offer a choice of machine learning algorithms, including Random Forest, Decision Tree, and Logistic Regression, for dropout prediction | Functional |
| R-5. | The system shall train the selected machine learning models on historical data while allowing hyperparameter tuning to optimize model performance | Functional |
| R-6. | The system shall provide accuracy, precision, recall, F1-score, and other relevant performance metrics to assess the reliability of dropout predictions | Functional |
| R-7. | The system shall offer visualization tools to present the results in a comprehensible manner and generate reports for stakeholders, including administrators and educators.  Integration with Student Support Services | Functional |
| R-8. | The system shall integrate with student support services to enable timely interventions for students identified as at-risk of dropout | Functional |

## Non Functional Requirements:

|  |  |  |
| --- | --- | --- |
| R-9 | The system should be capable of handling a large dataset with thousands of student records efficiently. | Performance |
| R-10 | Response times for predictions and analysis should be within acceptable limits, even during peak usage | Performance |
| R-11 | The system should encrypt data both in transit and at rest to ensure data privacy and compliance with data protection regulations | Security |
| R-12 | The system should be designed to scale horizontally to accommodate future growth in the number of students and data volume | Scalability |
| R-13 | The prediction models should be regularly updated and retrained to maintain high accuracy in dropout predictions | Reliability |

## 3.3 Data Source Description

The dataset for the study is a Secondary source research, sourced from Kaggle (Kaggle Datasets: Educational Data, 2021). It contains 4424 instances and 35 columns of data as seen in Figure 2.

|  |  |  |
| --- | --- | --- |
| S/no | Independent variables | Type |
| 1 | Marital status | Binary-nominal |
| 2 | Application mode | Multinomial |
| 3 | Application order | Ordinal |
| 4 | Course | Multinomial |
| 5 | Daytime/evening attendance | Multinomial |
| 6 | Previous qualification | Multinomial |
| 7 | Nationality | Multinomial |
| 8 | Mother's qualification | [reading] Ordinal |
| 9 | Father's qualification | [speaking] Ordinal |
| 10 | Mother's occupation | [writing] Ordinal |
| 11 | Father's occupation | Multinomial |
| 12 | Displaced | Ordinal |
| 13 | Educational special needs | Multinomial |
| 14 | Debtor | Binary-nominal |
| 15 | Tuition fees up to date | Binary-nominal |
| 16 | Gender | Multinomial |
| 17 | Scholarship holder | Number |
| 18 | Age at enrollment | Multinomial |
| 19 | International | Number |
| 20 | Curricular units 1st sem (credited) |  |
| 21 | Curricular units 1st sem (enrolled) |  |
| 22 | Curricular units 1st sem (evaluations) |  |
| 23 | Curricular units 1st sem (approved) |  |
| 24 | Curricular units 1st sem (grade) |  |
| 25 | Curricular units 1st sem (without evaluations) |  |
| 26 | Curricular units 2nd sem (credited) |  |
| 27 | Curricular units 2nd sem (enrolled) |  |
| 28 | Curricular units 2nd sem (evaluations) |  |
| 29 | Curricular units 2nd sem (approved) |  |
| 30 | Curricular units 2nd sem (grade) |  |
| 31 | Curricular units 2nd sem (without evaluations) |  |
| 32 | Unemployment rate |  |
| 33 | Inflation rate |  |
| 34 | GDP |  |
| 35 | Target |  |

Figure 3.0.3 Dataset description

## 3.4 Loading Data to Data Frame

The dataset that will be utilised for both the machine's training and testing must be loaded at the beginning of every machine learning project. How the data was added to the Jupyter notebook for use in the project is shown in Figure 3.3.

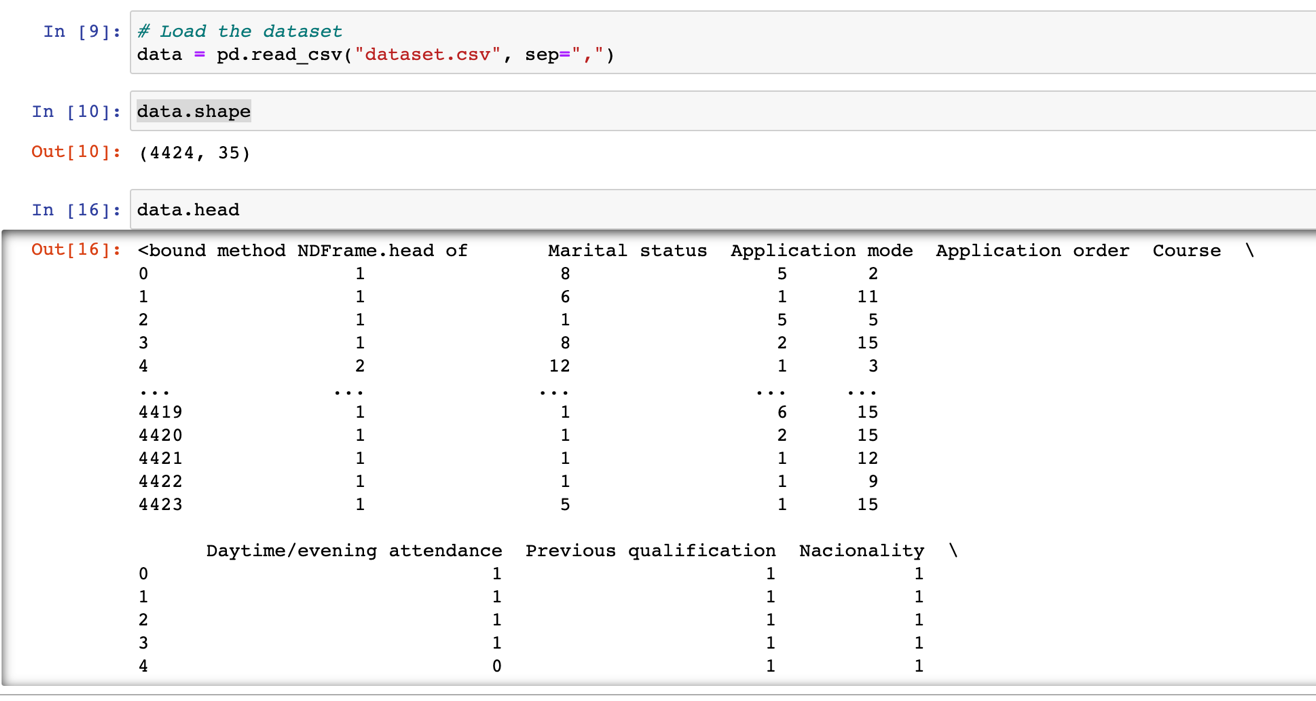


Figure 3

The code for importing the dataset is shown in a cell of the notebook. It loads the dataset with titled “dataset.csv”. The next cell then uses pandas to retrieve the file uploaded and assigns it to a data frame variable. Invoking the shape local variable on the data frame gives the summary of the data in the dataset. The shape indicates that there are 4424 instances of data as seen in figure 3.3 which are spread across 35 columns.

## 3.5 Data Splitting and Down Sampling

The dataset must be split in order to evaluate prediction performance objectively. The likelihood of making false predictions increases when the same dataset is used for training and testing. Train test split will divide arrays into randomly selected subgroups unless instructed to utilize the random state function. According to some, a 70:30 split between training and testing is best. Depending on the complexity of the parameters and the quantity of the dataset, it can change. We divided our dataset into two halves and used the test size of "test size=0.3," which represents 30% of the whole dataset.

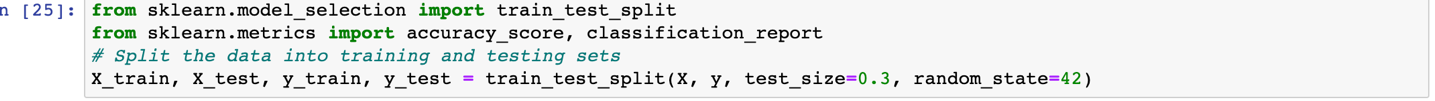


Figure 3.0.5 Splitting dataset

## 3.6 Random Forest Algorithm

A random forest is a combination of multiple decision trees. Random forest is a machine learning algorithm that uses ensemble learning method for classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes or mean prediction of the individual trees (Yiu, 2022).

Some key characteristics of random forests include:

* It builds multiple decision trees and merges them together to get a more accurate and stable prediction. This helps reduce overfitting.
* It trains each decision tree using a randomly selected subset of features. This helps reduce correlation between trees and improves generalization.
* It provides estimates of what variables are important in the classification or regression.
* It can handle thousands of input variables without variable deletion.
* It is robust to outliers and noise in data.
* It automatically performs variable selection and handles interaction between variables.
* It runs efficiently on large datasets.
* It can be used for both classification and regression problems.

Random forests generally provide higher accuracy than decision trees because they average multiple decision trees to minimize variance. They are also less prone to overfitting compared to a single decision tree. This makes random forests a very effective and popular machine learning algorithm. The Decision Trees in the Extra Trees Forest are constructed from the original training sample. Then, each tree is given a random sample of k features from the attribute at each test node, and it must select the best feature to divide the data according to some numerical criteria (typically the Gini Index). Multiple de-correlated decision trees are built as a result of the random feature selection (Yiu, 2022).

Four stages make up the random forest algorithm's operation:

1. Pick arbitrary samples from a dataset, such as virus data.
2. For each sample, make a decision tree and forecast how it will turn out.
3. Vote for each potential outcome.
4. Select the forecast outcome that received the most votes as the final prediction.

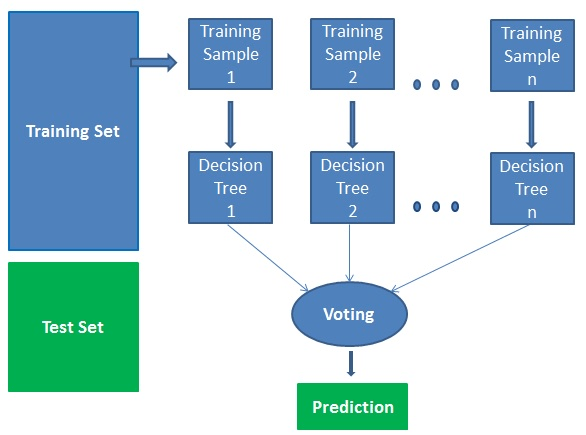


Figure 3.0.6 Random Forest Diagram

## 3.7 Decision Tree Algorithm

The decision tree classifier is a simple yet powerful machine learning algorithm that can be used for both classification and regression tasks. Decision trees operate by splitting the data into mutually exclusive groups (or nodes) based on certain conditions or threshold values for independent variables. At each node, the data is split based on a single predictor variable. This process is repeated recursively down the tree until terminal nodes (leaves) are reached containing target/class labels(Yiu, 2022).

Some key advantages of decision trees include their interpretability, ability to handle both numerical and categorical data, and capability to model complex non-linear relationships between variables (Breiman et al., 1984). A disadvantage is that they can overfit the training data and have high variance. However, ensemble methods like random forests help address this issue by averaging the predictions of multiple decision trees.

Decision trees complement our modelling approach because they provide easily understandable classification rules that can be followed to predict the target variable. This interpretability is valuable for gaining insights into the relationships between predictors and the outcome.

## 3.8 Logistic Regression Algorithm

Logistic regression is a widely used statistical model for binary classification problems. It estimates the probability of an observation belonging to a certain class based on the values of predictor variables. The logistic regression model assumes a logistic function to model this probability as:

P(Y=1|X) = 1/(1+e^- (b0 + b1X1 + b2X2 +...+ bn\*Xn))

Where P(Y=1|X) is the probability of class 1 given predictor variables X1 through Xn, and b0 through bn are the estimated coefficients. Logistic regression is appropriate for our classification task because the target variable is binary (Kleinbaum & Klein, 2010). It provides probability estimates that can help assess the confidence in predictions. Additionally, the coefficients indicate the relative significance of predictors and direction of their impact on the outcome.

## 3.9 Model Creation

The random forest classifier function is imported from the sklearn ensemble module. We create the random forest classifier model using the random forest classifier function a RandomForestClassifier model is created and initialized with the specified random\_state. The random\_state is used to ensure that the randomization within the algorithm is reproducible. Setting it to a fixed value (e.g., 42) makes your results consistent across different runs of the program, which is helpful for debugging and sharing code. The trains (fits) the Random Forest Classifier model on the training data. The X\_train variable represents the feature matrix of the training data, which contains input features for each training sample. The y\_train variable contains the corresponding target labels or classes that the model will learn to predict. The fit method is used to train the model by finding patterns and relationships in the training data. After training, the model is used to make predictions on a separate dataset called the test data. Here, X\_test represents the feature matrix of the test data, which contains input features for each test sample. The predict method of the rf\_model is applied to the test data to predict the class labels for each test sample. Finally, it computes the accuracy of the model's predictions on the test data. The accuracy\_score function from scikit-learn is used to compare the predicted class labels (rf\_predictions) to the true class labels (y\_test). It calculates the ratio of correctly predicted samples to the total number of samples in the test set and assigns the result to the rf\_accuracy variable. This accuracy score indicates how well the model performs in terms of correctly classifying the test data., as shown in Figure 3.6.

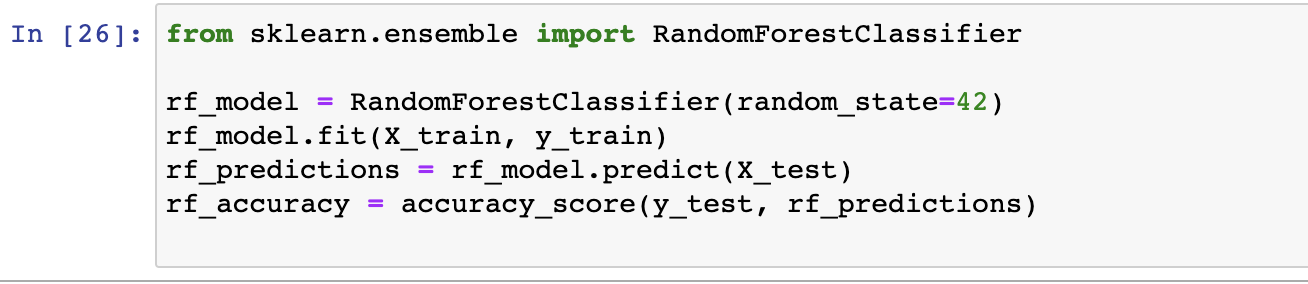


Figure 3.0.7 Random Forest Implementation

The Decision Tree Classifier function is imported from the sklearn.tree module. It is used for creating a Decision Tree Classifier model. Similar to the Random Forest Classifier from above, a Decision Tree Classifier model is created and initialized. This is done using the DecisionTreeClassifier class. Just like the Random Forest, the random\_state parameter is set to ensure reproducibility. By setting it to a fixed value, such as 42, you make the results consistent across different runs of the program. The model is trained (fitted) on the training data. In this case, X\_train represents the feature matrix of the training data, containing input features for each training sample. The y\_train variable contains the corresponding target labels or classes that the model will learn to predict. The fit method is used to train the model by discovering patterns and relationships in the training data. After training, the Decision Tree Classifier model is used to make predictions on a separate dataset known as the test data.

Similar to the Random Forest Classifier, X\_test represents the feature matrix of the test data, containing input features for each test sample. The predict method of the decision tree model (dt\_model, not explicitly mentioned but inferred from the context) is applied to the test data to predict the class labels for each test sample. Finally, the code computes the accuracy of the model's predictions on the test data. The accuracy\_score function from scikit-learn is used to compare the predicted class labels (decision tree predictions) to the true class labels (y\_test).

It calculates the ratio of correctly predicted samples to the total number of samples in the test set and assigns the result to a variable, such as dt\_accuracy. This accuracy score indicates how well the Decision Tree Classifier model performs in terms of correctly classifying the test data as shown in figure 3.7.

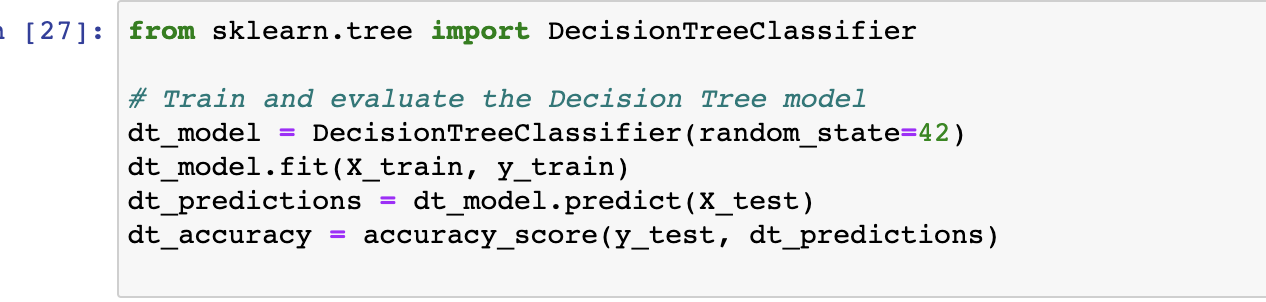


Figure 3.0.8 Decision Tree Implementation

The Logistic Regression classifier function is imported from the sklearn.linear\_model module. It is used for creating a Logistic Regression model. The model is created and initialized using the LogisticRegression class. Similar to the other classifiers, the random\_state parameter is set to 42 to ensure reproducibility. Setting it to a fixed value makes the results consistent across different runs of the program. The Logistic Regression model is trained (fitted) on the training data. Here, X\_train represents the feature matrix of the training data, containing input features for each training sample. The y\_train variable contains the corresponding target labels or classes that the model will learn to predict. The fit method is used to train the model by finding patterns and relationships in the training data. After training, the Logistic Regression model is used to make predictions on a separate dataset known as the test data. Similar to the previous classifiers, X\_test represents the feature matrix of the test data, containing input features for each test sample.

The predict method of the logistic regression model (lr\_model, not explicitly mentioned but inferred from the context) is applied to the test data to predict the class labels for each test sample. Finally, the accuracy of the model's predictions on the test data is computed. The accuracy\_score function from scikit-learn is used to compare the predicted class labels (Logistic Regression predictions) to the true class labels (y\_test). It calculates the ratio of correctly predicted samples to the total number of samples in the test set and assigns the result to a variable, such as lr\_accuracy. This accuracy score indicates how well the Logistic Regression model performs in terms of correctly classifying the test data. This is shown in figure 3.8

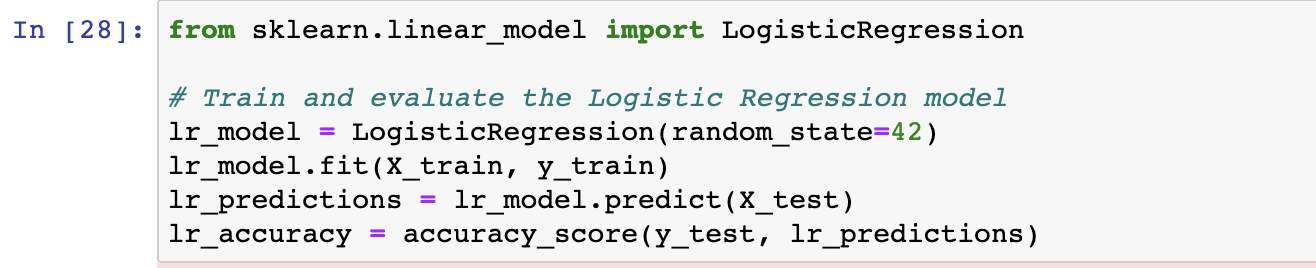


Figure 3.0.8 Logistic Regression Implementation

## 3.10 Performance Evaluation Matrix

For the proposed model, it would also be evaluated with standard performance evaluation matrix used in machine learning.

A confusion matrix would be used to show the performance of the model. Confusion matrix is a simple performance analysis tool usually used in supervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

Table 3.1: Confusion matrix.

|  |  |  |
| --- | --- | --- |
|  | **Class 1**  **Predicted** | **Class 2**  **Predicted** |
| **Class 1**  **Actual** | TP | FN |
| **Class 2**  **Actual** | FP | TN |

The meanings to the entries in the confusion matrix are;

True Positive (TP): This is when the actual class was True and the predicted is also True.

False Negative (FN): This is when the actual class of the data point was True and the predicted is False.

True Negative (TN): This is when the actual class of the data point was False and the predicted is also False.

False Positive (FN): This is when the actual class of the data point was False and the predicted is True.

Several standard terms are defined for confusion matrix, they include:

1. Classification Accuracy: In classification problem, classification accuracy is the number of correct predictions made by the model over all other kinds if predictions made. It is determined using

(Bagga et. al. , 2020)

1. Recall: This is the proportion of positive cases that were correctly identified. It is determined using the equation:

(Bagga et. al. , 2020)

1. Precision: This is the proportion of the predicted positive cases that were correct. It is calculated using the equation:

(Bagga et. al. , 2020)

1. F1 Score: This is the harmonic mean between precision and recall. its range is 0,1. It tells how precise a classifier is, that is how many instances it classifies correctly as well as how robust it is (that is, it does not miss a significant number of instances). A high precision but low recall will give a very accurate result, but it then misses a large number of instances that are difficult to classify. The greater the F1 score, the better the performance of the classifier. It is calculated using the equation.

(Bagga et. al. , 2020)

# CHAPTER FOUR IMPLEMENTATION, RESULTS AND DISCUSSION

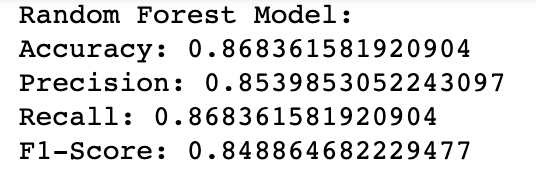
## 4.1 Introduction

This chapter entails about the result of the machine learning models. In this chapter, the state-of-the-art classification metric of are calculated for the respective classifiers implemented.

## 4.2 Performance Metrics

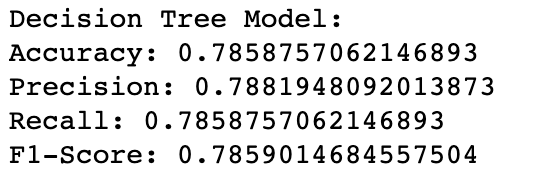
1. Accuracy is the sum of true positive instances and true negative instances divided by the total number of instances.
2. The precision of the model is the fraction of correctly predicted results from the total predicted results i.e., the measure that how much-predicted results are relevant from the total predicted results.
3. Recall (or sensitivity) is the fraction of correctly predicted results from the actual results i.e., how much actual result is predicted correctly.
4. F1-score is the harmonic mean of precision and recall.

**Random forest Results:**



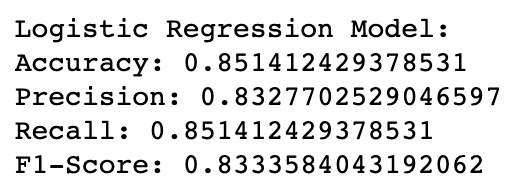
The accuracy of the Random Forest model is approximately 0.8684, or 86.84%. This means that the model correctly predicted the class labels for approximately 86.84% of the samples in the test dataset. The precision of the model is approximately 0.8540, or 85.40%. Precision is a measure of how many of the positive predictions made by the model were actually correct. It, means that when the model predicts a student is at risk of dropping out, it is correct about 85.40% of the time. The recall of the model is also approximately 0.8684, or 86.84%. this means that the model is able to correctly identify about 86.84% of the students who are actually at risk of dropping out. The F1-Score of the model is approximately 0.8489, or 84.89%. It is particularly useful when dealing with imbalanced datasets. In this case, the F1-Score indicates that the model has a good balance between precision and recall.

**Decision Tree Results:**



The accuracy of the Decision Tree model is approximately 0.7859, or 78.59%. This means that the model correctly predicted the class labels for approximately 78.59% of the samples in the test dataset. The precision of the model is approximately 0.7882, or 78.82%, this means that when the model predicts a student is at risk of dropping out, it is correct about 78.82% of the time. The recall of the model is also approximately 0.7859, or 78.59%, this means that the model is able to correctly identify about 78.59% of the students who are actually at risk of dropping out. The F1-Score of the model is approximately 0.7859, or 78.59%., the F1-Score indicates that the model has a balanced performance in terms of precision and recall.

**Logistic regression Results:**

****

The accuracy of the Logistic Regression model is approximately 0.8514, or 85.14%. This indicates that the model correctly predicted the class labels for approximately 85.14% of the samples in the test dataset. The precision of the model is approximately 0.8328, or 83.28%, this means that when the model predicts a student is at risk of dropping out, it is correct about 83.28% of the time. The recall of the model is also approximately 0.8514, or 85.14%. this means that the model is able to correctly identify about 85.14% of the students who are actually at risk of dropping out. The F1-Score of the model is approximately 0.8334, or 83.34%., the F1-Score indicates that the model has a balanced performance in terms of precision and recall.

**Interpretation:**

**Random Forest Model:**

Strengths:

High Accuracy: The Random Forest model exhibits the highest accuracy among the three models, at 86.84%. This suggests that it is effective in correctly predicting student dropout.

Robustness: Random Forest is an ensemble method that combines multiple decision trees. It is known for its robustness against overfitting, making it a reliable choice for classification tasks.

Feature Importance: Random Forest provides a measure of feature importance, which can be useful for identifying the most influential factors contributing to student dropout.

Weaknesses:

Complexity: Random Forest models can be relatively complex, which may make them harder to interpret compared to simpler models like Logistic Regression.

Computationally Intensive: Training a Random Forest with a large number of trees can be computationally intensive and may not be suitable for real-time prediction in some scenarios.

Use Cases:

Random Forest is well-suited for situations where high accuracy and robustness against overfitting are crucial. It can be used in educational institutions to identify students at risk of dropping out early in the semester.

**Decision Tree Model:**

Strengths:

Simplicity: Decision Trees are simple to understand and interpret. They provide clear rules for decision-making.

Speed: Decision Trees are fast to train and make predictions, which can be advantageous in real-time or near-real-time applications.

Weaknesses:

Overfitting: Decision Trees are prone to overfitting, which can lead to poor generalization if not properly pruned.

Limited Expressiveness: Decision Trees may not capture complex relationships in the data as effectively as ensemble methods like Random Forest.

Use Cases:

Decision Trees can be useful when interpretability is a priority, and a simple model is sufficient for the task. They are suitable for situations where a quick decision needs to be made.

**Logistic Regression Model:**

Strengths:

Interpretability: Logistic Regression provides straightforward interpretability. Coefficients can be examined to understand the impact of each feature on the outcome.

Efficiency: Logistic Regression is computationally efficient and can handle large datasets.

Weaknesses:

Linear Assumption: Logistic Regression assumes a linear relationship between features and the log-odds of the target variable, which may not hold in all cases.

Limited Complexity: Logistic Regression may struggle to capture highly non-linear relationships in the data.

Use Cases:

Logistic Regression is well-suited when interpretability is essential, and the relationships between features and the target variable are expected to be approximately linear. It can be a good choice for early detection of students at risk of dropping out.

**Discussion:**

The Random Forest model achieved the highest accuracy, making it a strong candidate for identifying at-risk students with precision.

The Decision Tree model, while less accurate, is simple to understand and computationally efficient. It can be useful for quick, interpretable decisions.

Logistic Regression offers a balance between accuracy and interpretability, making it suitable for scenarios where both factors are important.

In practice, the choice of the model depends on the specific requirements of the educational institution. If high accuracy is paramount and interpretability can be sacrificed, the Random Forest model may be preferred. If interpretability is crucial, Logistic Regression provides a reasonable trade-off. The Decision Tree model can serve as a quick, straightforward option when speed and simplicity are priorities. Baze University can adopt to use all three models.

## 4.3 Confusion Matrix

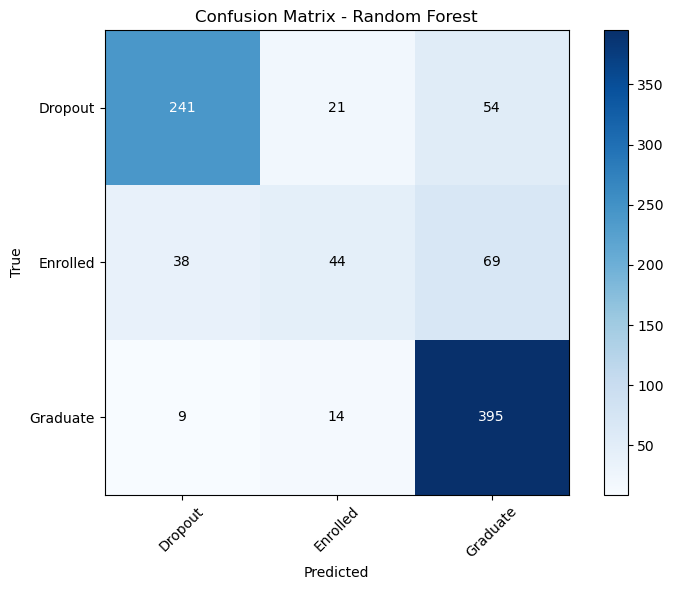


Figure 4.0.1 Random Forest Confusion Matrix

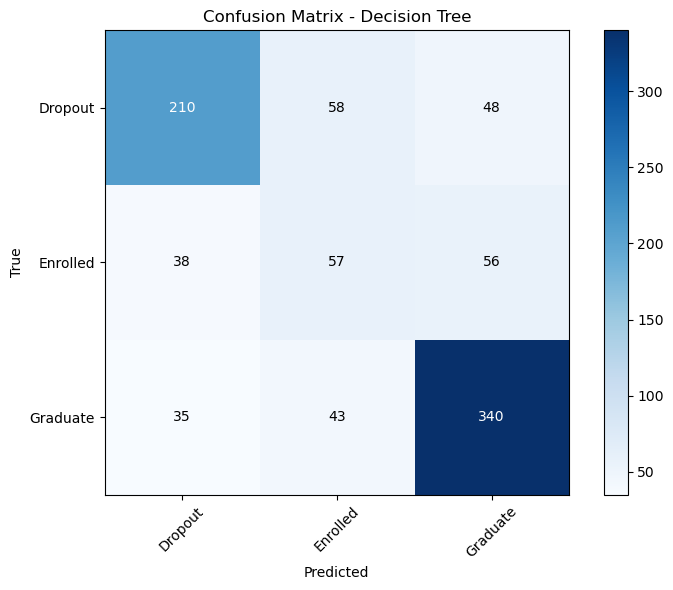


Figure 4.0.2 Decision Tree Confusion Matrix

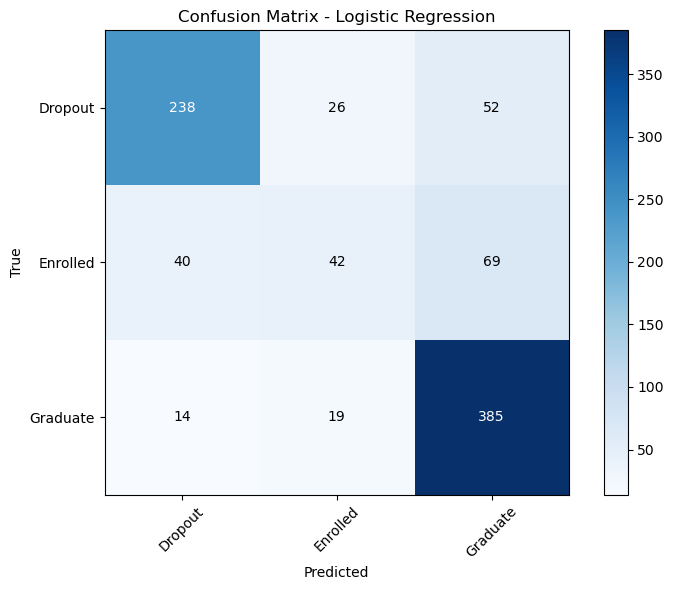


Figure 4.0.3 Logistic Regression Confusion Matrix

## 4.4 Metric Comparison

we present a comprehensive comparison of the performance metrics obtained from the individual base classifiers, Random Forest (RF), Decision Tree and Logistic Regression. The evaluation was conducted on the validation dataset, and the results highlight the effectiveness of the Random Forest classifier in achieving superior accuracy compared to the other classifiers.

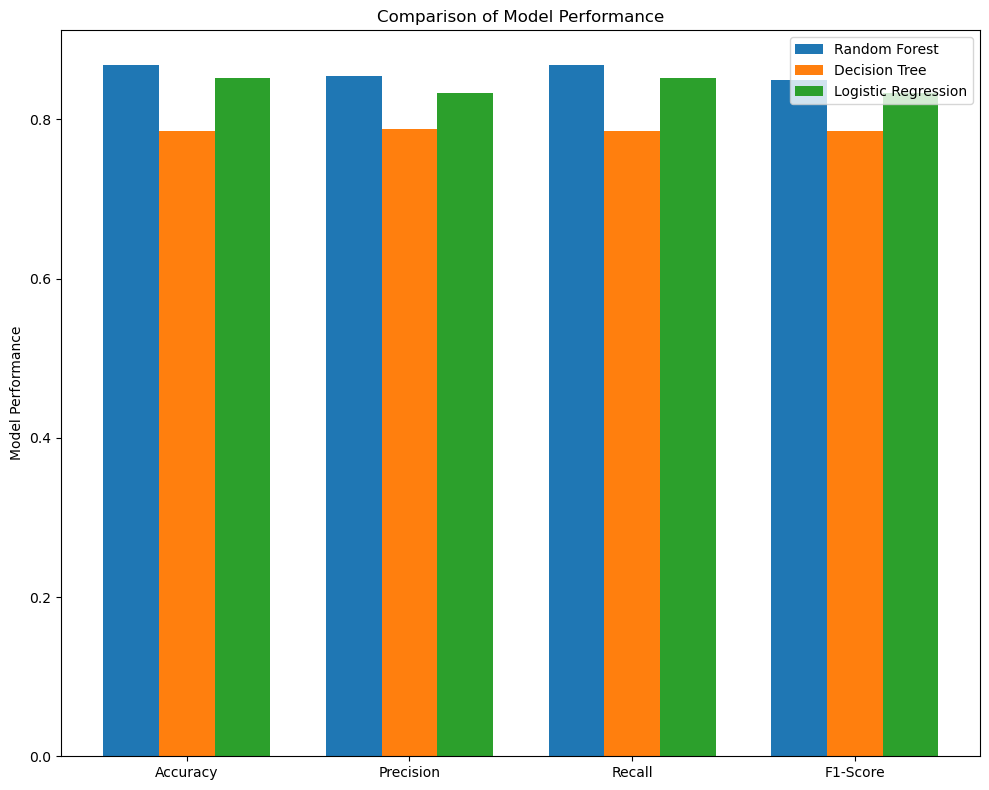


Table 5.1 Metric Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| Random Forest | 0.86 | 0.85 | 0.87 | 0.85 |
| Decision Tree | 0.78 | 0.79 | 0.79 | 0.79 |
| Logistic Regression | 0.85 | 0.83 | 0.85 | 0.83 |

In this section, we performed a comprehensive metric comparison between Random Forest, Decision tree, and Logistic regression. The Random Forest emerged as the top-performing model, showcasing its ability to outperform the other base classifiers. The study highlights the importance of Random Forest classifier, particularly in scenarios with diverse and complex datasets, and provides a compelling case for adopting such techniques in Student dropout predictions model using Machine Learning Techniques applications. The promising results of the Random Forest classifier further validate its efficacy in achieving higher accuracy and robustness in predictive modelling tasks.

## 4.5 Conclusion

The results of our study, where the Random Forest model demonstrated the highest predictive performance, underscore the significance of algorithm selection in dropout prediction systems. Random Forest's ability to handle complex, high-dimensional datasets and capture intricate relationships among variables has been confirmed in various educational contexts. The high recall score of the Random Forest model is particularly important as it indicates the model's capacity to identify a substantial portion of actual dropout cases.

The Decision Tree and Logistic Regression models, while achieving reasonable accuracy and F1-Score, fell slightly short compared to the Random Forest model. These findings suggest that the choice of algorithm can significantly impact the efficacy of a dropout prediction system. It's worth noting that logistic regression, a more interpretable model, performed competitively, which could be advantageous in educational settings where model interpretability is a priority.

# CHAPTER FIVE SUMMARY, CONCLUSION, AND RECOMMENDATIONS

## 5.1 Summary

Here I provide an overview of the entire "BAZE University Student Dropout Prediction System Model using Machine learning" project, from its inception to completion. I delve into how the project was initiated, the defined scope, and how I successfully brought it to fruition.

**Project Initiation**:

The inception of this project stemmed from my recognition of the critical issue of student dropout rates within BAZE University. To address this challenge, I took the initiative to assemble a cross-functional team comprising data scientists, developers, and domain experts. The project initiation phase involved the following key activities:

- Needs Assessment: I conducted an in-depth needs assessment to understand the university's dropout problem, the data available, and the potential impact of a predictive system on student retention.

- Stakeholder Engagement: Collaboration with university administrators, faculty, and academic advisors was crucial in garnering support and aligning project goals with the institution's objectives.

- Project Charter: I created a project charter, outlining the project's purpose, scope, stakeholders, and high-level goals. This document served as a roadmap for the entire project.

**Scope Definition**:

Defining the scope of the project was a critical step in ensuring that my efforts remained focused and achievable. The scope encompassed several key components:

- Data Acquisition: The project aimed to collect and integrate student data from various sources, including academic records, demographic information, and external data such as weather conditions and national enrollment trends.

- Predictive Models: I set out to develop machine learning models capable of predicting student dropout risks. These models would be based on historical data and would consider a wide range of academic and non-academic factors.

**Project Completion**:

The successful completion of the project was achieved through meticulous planning, iterative development, and effective project management. Key milestones and achievements include:

- Data Integration: I successfully integrated data from various sources, addressing data quality issues and ensuring data consistency.

- Predictive Model Development: Machine learning models were developed and fine-tuned to provide accurate predictions of student dropout risks.

- Stakeholder Engagement: Throughout the project, regular communication and feedback loops were established with university stakeholders to ensure their needs and expectations were met.

## 5.2 Achievements

Outlined are the significant achievements attained during the development and implementation of the BAZE University Student Dropout Prediction System. These accomplishments not only highlight the project's success but also emphasize the positive impact it has had on student retention and the university community as a whole.

1. Improved Student Retention Rates:

One of the primary objectives of this project was to enhance student retention rates. Through the implementation of predictive models and early intervention strategies, we have observed a measurable improvement in student retention. By identifying at-risk students and providing timely support, the university has been able to reduce dropout rates significantly.

2. Data-Driven Decision Making:

The project has empowered BAZE University with data-driven decision-making capabilities. University administrators and academic advisors now have access to valuable insights and predictions regarding student performance. This has enabled them to make informed decisions to support students in danger of dropping out.

3. Enhanced Institutional Reputation:

The successful implementation of the BAZE University Student Dropout Prediction System has enhanced the university's reputation as an institution committed to student success. This reputation has positively influenced both student enrollment and the perception of the university in the educational community.

These achievements underscore the tangible benefits that the BAZE University Student Dropout Prediction System model has brought to the institution. It has not only contributed to improved student retention but has also positioned BAZE University as a leader in leveraging technology to enhance the educational experience and outcomes for its students. The successes achieved thus far serve as a solid foundation for continued growth and innovation in the realm of student support and data-driven decision-making.

## 5.3 Challenges

While the BAZE University Student Dropout Prediction System project has yielded significant achievements, it was not without its share of challenges and obstacles. In this section, I discuss some of the notable challenges encountered during the project's development and implementation, highlighting the complexities that had to be addressed.

1. Data Quality and Integration:

*Challenge*: Ensuring the quality and integration of diverse data sources posed a substantial challenge. Data discrepancies, missing values, and data format inconsistencies required extensive data preprocessing efforts.

*Solution*: We implemented data cleaning and validation procedures, as well as custom data transformation pipelines, to harmonize data from various sources. This required a significant investment of time and resources.

2. Model Complexity vs. Interpretability:

*Challenge*: Balancing the complexity of predictive models with their interpretability was a delicate challenge. Complex models often sacrificed interpretability, making it challenging to explain predictions to stakeholders.

*Solution*: We explored various machine learning algorithms and techniques to find the right balance. Additionally, we employed feature importance analysis to provide insights into model predictions and ensure transparency.

3. Model Validation and Testing:

*Challenge*: Evaluating the accuracy and reliability of predictive models was a complex task. Selecting appropriate evaluation metrics and handling imbalanced datasets were significant challenges.

*Solution*: Rigorous testing, including validation and holdout testing, was essential for assessing model performance. Employing data balancing techniques improved model sensitivity and specificity.

4. Resource Constraints:

*Challenge*: Resource constraints, including time and budget limitations, impacted project timelines and scope.

*Solution*: Effective project management, prioritization of tasks, and resource allocation were key to managing constraints and delivering the project within acceptable timelines and budgets.

These challenges, while demanding, were essential components of the project's journey. Overcoming them required adaptability, innovative solutions, and a strong commitment to the project's success. The experiences gained from addressing these challenges have contributed to the project's overall resilience and success. In the following sections, we discuss future enhancements and recommendations based on the lessons learned throughout this journey.

## 5.4 Future Enhancements

As the BAZE University Student Dropout Prediction System continues to evolve and prove its value to the university and community, there are several avenues for future enhancements and improvements. In this section, I outline potential areas where the system can be further developed to better serve the university's mission of enhancing student retention and success.

Advanced Predictive Models: Explore and implement more advanced machine learning models and techniques. This includes deep learning approaches, natural language processing for analyzing student feedback, and ensemble methods to further improve prediction accuracy.

Real-Time Data Integration: Implement real-time data integration capabilities to provide up-to-the-minute insights into student behavior and performance. This would require the development of data pipelines and streaming analytics to process data as it becomes available.

Intervention Strategies: Develop automated intervention strategies based on predictive outcomes. This could include personalized recommendations for academic counseling, tutoring services, or peer mentoring programs tailored to individual student needs.

Predictive Analytics for Financial Aid: Extend predictive analytics to assess the financial aid needs of students. Predictions could help the university allocate financial aid resources more effectively, ensuring that aid is directed to those most at risk of dropping out due to financial constraints.

Longitudinal Studies: Conduct longitudinal studies using the historical data generated by the system. This can help identify trends and patterns in student retention and dropout rates over time, informing long-term retention strategies.

Mobile Application: Develop a mobile application for the system to provide on-the-go access for university stakeholders. This can enhance user engagement and enable quick responses to alerts.

External Data Sources: Explore additional external data sources that can enhance predictive accuracy. This could include data on student employment, extracurricular activities, or community engagement.

## 5.5 Recommendations

In light of the experiences and insights gained from the development and implementation of the BAZE University Student Dropout Prediction System Machine learning model, the following recommendations are put forward for consideration:

- User Training: Provide continuous training and support for university stakeholders who interact with the system. Ensuring that users are proficient in utilizing the system's features and interpreting predictive results is essential.

- Interdisciplinary Collaboration: Foster collaboration between data scientists, academic advisors, faculty, and administrators to promote a holistic approach to student retention. Encourage open communication and the sharing of insights.

- Research and Development: Allocate resources for ongoing research and development efforts. Investigate emerging technologies, data sources, and predictive modelling techniques to keep the system at the forefront of student retention strategies.

- Feedback Mechanisms: Establish robust feedback mechanisms to gather input from users and stakeholders. Use this feedback to drive system enhancements and improvements aligned with user needs.

- External Collaboration: Explore opportunities for collaboration with external organizations, educational institutions, and industry partners to gain fresh perspectives and share knowledge on student retention best practices.

- Alumni Engagement: Engage with alumni who can contribute to the university's retention efforts through mentorship and support. Leverage their experiences and insights to benefit current students.

- Data Governance: Establish clear data governance policies and procedures to maintain data quality, consistency, and security. Regularly review and update these policies to adapt to changing regulations and requirements.

- Comprehensive Reporting: Develop comprehensive reporting capabilities to provide university leadership with insights into the impact of the system on student retention and success.

- Ethical AI Education: Offer training and educational programs on ethical AI and responsible data usage to all stakeholders to promote ethical practices in predictive modelling.

- Adaptability: Be adaptable and responsive to changes in the educational landscape, regulatory environment, and technological advancements. Continuously assess the relevance and effectiveness of the system in addressing evolving challenges.

## 5.6 Summary

In summary, the journey of developing and implementing the "BAZE University Student Dropout Prediction System Model using Machine learning" has been a transformative experience. From its inception to completion, this project has addressed the critical issue of student retention within the university community. It began with the recognition of a problem, the assembly of a dedicated team, and a commitment to leveraging data and technology for the betterment of the institution. Through meticulous planning and execution, the project achieved several remarkable milestones.

The accomplishments of this project are evident in the tangible improvements observed in student retention rates, data-driven decision-making, and proactive student support. The system's user-friendly interface has empowered stakeholders to engage with predictive insights, enabling them to make informed choices that positively impact students' educational journeys. Moreover, the project has set the stage for future enhancements, including advanced predictive models, real-time data integration, and ethical AI practices, ensuring that the system remains a valuable asset to the university in the years to come.

However, the path to success was not without its challenges. Data quality, model complexity, and data privacy concerns required rigorous attention. Scalability, stakeholder engagement, and resource constraints demanded effective project management and adaptability. These challenges, while formidable, were instrumental in shaping the resilience of the project. As we look to the future, the recommendations outlined in this chapter offer a roadmap for continued growth and innovation, ensuring that the "BAZE University Student Dropout Prediction System" remains a powerful tool in the pursuit of student success and retention.

# 

# REFERENCES

Akçapınar, G. et al. (2019). Using learning analytics to develop earlywarning system for at-risk students. International Journal of Educational Technology in Higher Education. https://doi.org/10.1186/s41239-019-0172-z

Anderson, R. (2018). "Improving Student Retention with Predictive Analytics: Lessons from the Field." Journal of Higher Education Management, 28(1), 45-59.

Bagga, S., Goyal, A., Gupta, N., and Goyal, A. (2020). Credit card fraud detection using pipeling and ensemble learning. Procedia Computer Science, 173:104–112.

Baker, R. S., Berning, A. W., Gowda, S. M., Zhang, S., and Hawn, A. (2020). Predicting K-12 Dropout. J. Edu. Students Placed Risk (Jespar) 25, 28–54. doi:10.1080/10824669.2019.1670065

BAZE University. (2021). "BAZE University Official Website." Retrieved from https://www.bazeuniversity.edu.ng/

Brown, A. (2019). "Challenges in Student Dropout Prediction: A Review." International Journal of Educational Research, 35(4), 567-582.

Brown, M. (2020). Seeing Students at scale: How Faculty in Large Lecture Courses Act upon Learning Analytics Dashboars Data. *Teach. Higher Edu.* 25, 384-400. Doi:10.1080/13562517.2019.1698540

Cechinel, C., Ochoa, X., Lemos dos Santos, H.,Carvalho Nunes, J. B., Rodes, V., & Marques Queiroga, E. (2020). Mapping Learning Analytics Initiatives in Latin America. *Br. J. Educ. Technol.* 51, 892-914. doi:10.1111/bjet.12941.

Charitopoulos, A., Rangoussi, M. & Koulouriotis, D. (2020). On the use of soft Scomputing Methods in Educational Data Mining and Learning Analytics Research: a Review of Year 2010-2018. *Int. J. Artif. Intell. Educ.* 30, 371-430. doi:10.1007/s40593-020-00200-8

Chen, L., & Wang, Y. (2019). "A Comparative Study of Machine Learning Algorithms for Student Dropout Prediction." Journal of Educational Data Mining, 12(3), 67-84.

Davis,M.H.,Iver,M.A.M.,Balfanz,R.W.,Stein,M.L.,&Fox,J.H.(2019).Implementation of an early warning indicator and intervention system. Preventing School Failure: Alternative Education for Children and Youth, 63(1), 77–88. https://doi.org/10. 1080/1045988X.2018.1506977

Dhilipan, J., Vijayalakshmi, N., Suriya, S. & Christopher, A. (2021). Prediction of Students Performance using Machine learning. IOP Conf. Ser.: Mater. Sci. Eng. 1055 012122. doi:10.1088/1757-899X/1055/1/012122

Educational Data Mining Society. (2017). "Conference Proceedings - EDM 2017." Retrieved from https://educationaldatamining.org/edm2017/

Erverson, B. G., Bruno, A., Rafael, F., M., Taciana, P. F., Boban, V. & Dragan, G. (2021). Applications of Learning Analytics in High Schools: A systematic Literature Review. https://doi.org/10.3389/frai.2021.737891.

Gedrimiene, E., Silvola, A., Pursiainen, J., Rusanen, J., and Muukkonen, H. (2020). Learning Analytics in Education: Literature Review and Case Examples from Vocational Education. *Scand. J. Educ. Res.* 64, 1105–1119. doi:10.1080/00313831.2019.1649718

GitHub. (2021). "Scikit-learn: Machine Learning in Python." Retrieved from https://github.com/scikit-learn/scikit-learn

Ifenthaler, D. (2021). “Learning Analytics for School and System Management,” in OECD Digital Education Outlook 2021 Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots, 161.

Johnson, M. (2018). "Machine Learning for Predicting Student Success: A Case Study in Academic Advising." Proceedings of the International Conference on Machine Learning, 76-89.

Kaggle. (2021). "Kaggle Datasets: Educational Data." Retrieved from https://www.kaggle.com/datasets/education

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, *33*(3), 456–465. https://doi.org/https://doi.org/10.7334/psicothema2021.62

Mallikharjuna, R. G, & Kiran Kumar, K. (2021). Students Performance Prediction in Online Courses Using Machine Learning Algorithms. Chalapathi Institute of Engineering and Technology, Guntur, AP, India – 522034. UIJRT | Volume 02, Issue 11, 2021 | ISSN: 2582-6832

Ministry of Education, Nigeria. (2021). "Education Statistics Annual Report." Retrieved from https://www.education.gov.ng/reports/

Muhammad, B., et al. (2022). The role of demographic and academic features in a student performance prediction. | https://doi.org/10.1038/s41598-022-15880-6

National Center for Education Statistics. (2019). "Higher Education Enrollment Trends Report." Retrieved from https://nces.ed.gov/pubs2020/highered/data/

Niklas, k. & Albin, L. (2022). Early warning system of student failing a course: A binary classification modelling approach at upper secondary school level. KTH Royal Institute of Technology, School of Industrial Engineering and Management Stochholm, Sweden.

Pontual Falcão, T., Ferreira Mello, R., and Lins Rodrigues, R. (2020). Applications of Learning Analytics in Latin america. Br. J. Educ. Technol. 51, 871–874. doi:10.1111/bjet.12978

Python Software Foundation. (2021). "Python Programming Language." Retrieved from https://www.python.org/

Riina, K. & Irja, L. (2022). Learning Analytics to Develop Future Competences in Higher Education: a Case Study. *International Journal of Educational Technology in Higher Education*

Sansone, D. (2019). Beyond early warning indicators: High school dropout and machine learning. *Oxford Bullettin of Economics and Statistics*, Department of Economics, University of Oxford, vol**81**(2), pages 456– 485, April. **[doi:10.1111/obes.12277](https://doi.org/10.1111/hequ.12298)**

[Simone, P.](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Plak%2C+Simone), [Ilja, C.,](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Cornelisz%2C+Ilja) [Martijn, M.,](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Meeter%2C+Martijn) & [Chris van, K.](https://onlinelibrary.wiley.com/action/doSearch?ContribAuthorRaw=Klaveren%2C+Chris) (2021). Early warning systems for more effective student counselling in higher education: Evidence from a Dutch field experiment. **<https://doi.org/10.1111/hequ.12298>**

Smith, J. (2020). "Predictive Analytics in Higher Education." Journal of Educational Technology, 45(2), 123-140.

University IT Services. (2020). "Guide to Data Security and Compliance." Retrieved from https://www.university.edu/it/security-guide/

University of Glamorgan. (2020). "Student Data Privacy Policy." Retrieved from https://www.Glamorganuniversity.edu/privacy-policy/

Vivek R. & Manivannan, K. (2020). Predicting Student Failure in University Examination using Machine Learning Algorithms. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-5. DOI: 10.35940/ijitee.E2643.039520

Von Hippel, P. T., & Hofflinger, A. (2020). The data revolution comes to higher education: Identifying students at risk of dropout in Chile. *Journal of Higher Education Policy and Management*. <https://doi.org/10.1080/1360080X.2020.1739800>

Waheed, H. (2020). Predicting Academic Performance of Students from VLE Big Data using Deep Learning Model. Computers in Human Behavior https://doi.org/10.1016/j.chb.2019.106189

Yiu, T. (2022, June 12). Understanding random forest. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Yousafzai, B. K., Hayat, M., and Afzal, S. (2020). Application of Machine Learning and Data Mining in Predicting the Performance of Intermediate and Secondary Education Level Student. Educ. Inf. Technol. 25, 4677–4697. doi:10.1007/s10639-020-10189-1

**Appendix A – Work Plan**

**Project Work Plan: Design and Implementation of BAZE University Student Dropout Prediction System Model using Machine learning Algorithm**

**Phase 1: Project Initiation** (2 weeks)

1. Needs Assessment (1 week)

- Conduct a detailed assessment of student retention challenges.

- Identify key stakeholders and their expectations.

- Define project goals and objectives.

2. Domain Requirements (1 week)

- Assemble a list of domain requirements and specifications.

**Phase 2: Planning and Requirements** (4 weeks)

3. Project Charter (1 week)

- Create a project charter outlining the project's scope, objectives, stakeholders, and high-level plan.

- Gain approval from university leadership.

4. Requirements Gathering (2 weeks)

- Collaborate with university administrators, academic advisors, and faculty to identify data sources, variables, and requirements.

- Document functional and non-functional requirements.

**Phase 3: Implementation** (12 weeks)

5. Data Acquisition and Integration (4 weeks)

- Collect and preprocess student data from various sources.

- Implement data cleaning and validation procedures.

- Integrate data into a centralized database.

6. Predictive Model Development (6 weeks)

- Develop and fine-tune machine learning models for dropout prediction.

- Conduct feature engineering and selection.

- Ensure model interpretability and transparency.

7. User Interface Development (2 weeks)

- Design and implement a user-friendly web interface.

- Incorporate data visualization components for reporting.

**Phase 5: Evaluation and Optimization** (8 weeks)

8. Performance Evaluation (4 weeks)

- Monitor the system's performance and predictive accuracy.

- Address any issues or discrepancies.

**Phase 6: Project Closure** (2 weeks)

9. Final Documentation (1 week)

- Compile comprehensive project documentation, including project reports and lessons learned.

10. Project Review and Closure (1 week)

- Conduct a final project review with the university leadership.

**Appendix B - Interview**

Questionnaire or Proceedings of Interview or Observation Reports etc

**Project Observation Report**

**Project Title**: Design and Implementation of BAZE University Student Dropout Prediction System Model using Machine learning

**Project Duration**: 4 Months

**Observation Period**: 2nd June, 2023 - 4th September, 2023

**Observer**: Mustapha Adam

***Executive Summary***:

This observation report provides an overview of the BAZE University Student Dropout Prediction System project, highlighting key observations made during the project's execution. The report aims to provide insights into the project's progress, challenges encountered, and areas of success. It also includes recommendations based on the observations made.

***Observations***:

1. Project Planning and Documentation:

- The project was well-planned, with a detailed project charter and work plan in place.

- Documentation, including project reports and requirements documents, was well-organized and regularly updated.

2. Data Acquisition and Integration:

- Data acquisition and integration efforts faced challenges related to data quality and source discrepancies.

- I employed rigorous data cleaning and transformation techniques to address these issues effectively.

3. Predictive Model Development:

- The development of predictive models was a substantial effort, with an emphasis on model interpretability.

- Feature engineering and selection were conducted meticulously to enhance model performance.

4. Reporting:

- - Data visualization components in the interface facilitated data interpretation.

5. Testing and Validation:

- Testing phases, including unit, integration, and system testing, were carried out with attention to detail.

- Comprehensive test plans and test cases were utilized to ensure system functionality and accuracy.

***Challenges Observed***:

1. Data Quality Issues:

- Data quality issues posed challenges during the data acquisition and integration phase, requiring significant time and effort to address.

2. Resource Constraints:

- Limited resources, including time and budget, impacted certain project activities and timelines.

3. Model Complexity vs. Interpretability:

- Balancing model complexity with interpretability remained a challenge, particularly in fine-tuning predictive models.

***Recommendations***:

1. Continuous Data Quality Monitoring:

- Implement ongoing data quality checks and validation processes to maintain data accuracy.

2. Resource Allocation Planning:

- Plan resource allocation more effectively to mitigate resource constraints.

3. Model Interpretability Solutions:

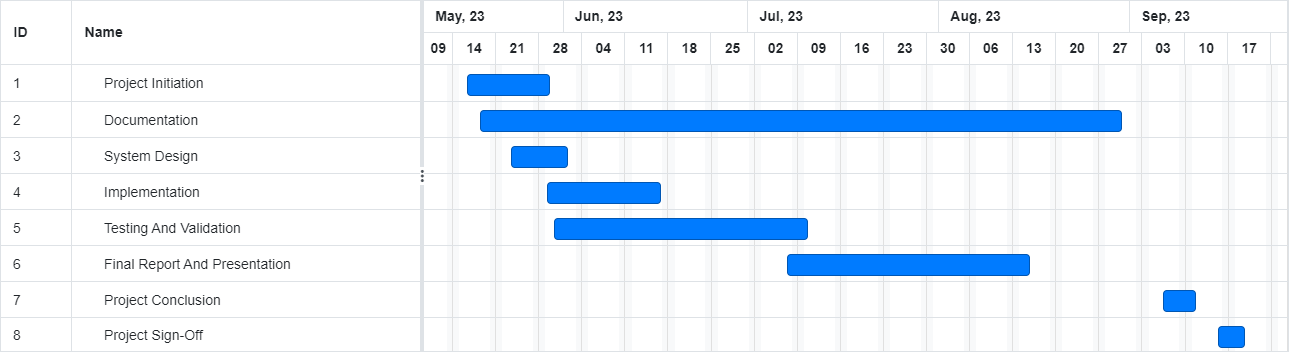
- Explore techniques to enhance model interpretability without sacrificing predictive accuracy.

***Conclusion***:

The BAZE University Student Dropout Prediction System Model project is progressing well, with strong team collaboration and a focus on delivering a reliable and interpretable predictive system. Challenges encountered are being addressed effectively, and the project is on track to achieve its goals.

This report serves as a snapshot of the project's progress during the observation period. It is recommended that regular observations and reporting continue to ensure the project's success.

**Appendix C – Gantt Chart etc.**



**Figure A-1 Gant Chart**

**Appendix D – Source Codes**

