

**DEVELOPMENT OF A PREDICTIVE ANALYTICS SYSTEM FOR STUDENTS’ PERFORMANCE**

**BY**

**ADEGBENRO, AFEEZ ADESHOLA**

**21/52HL155**

**A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A BACHELOR OF SCIENCE (B.Sc.) DEGREE IN INFORMATION TECHNOLOGY**

**DECEMBER, 2024**

# Certification Page

This is to certify that this project work was carried out by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ with matriculation number \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ in the department of Information and Communication Science, University of Ilorin, Ilorin Nigeria.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Name) Signature and Date

Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Name) Signature and Date

Head of Department

# Dedication

I, Adegbenro Afeez Adeshola, declare that the work carried out and reported on by this document and the contents of this document, titled **Predictive Analytics for Students Performance**, and accompanying artefacts, except for where attributed to other sources by way of citations within the body of this document and listed under the References section of this document, is formed of my own original work, which has never before been produced for any other purpose, and has been supervised by Prof[…].

All information in this document and accompanying artefacts has been obtained and presented in accordance with the relevant ethical conduct and academic rules.

**Signature**: Adegbenro Afeez Adeshola

# Acknowledgment

Firstly, I would like to tender my profound appreciation to God for being there and helping all through the course of my final year project and academic journey.

Table Of Contents

[Certification Page 2](#_Toc194838267)

[Dedication 3](#_Toc194838268)

[Acknowledgment 4](#_Toc194838269)

[Abbreviations and Acronyms 9](#_Toc194838270)

[Abstract 10](#_Toc194838271)

[INTRODUCTION 11](#_Toc194838272)

[1.1 Background to the Study 11](#_Toc194838273)

[1.2 Statement of the Problem 13](#_Toc194838274)

[1.3 Aim and Objectives 13](#_Toc194838275)

[1.4 Research Questions 14](#_Toc194838276)

[1.5 Significance of The Study 15](#_Toc194838277)

[1.6 Scope of The Study 16](#_Toc194838278)

[1.7 Limitations of The Study 16](#_Toc194838279)

[LITERATURE REVIEW 18](#_Toc194838280)

[2.1 Introduction to Educational Data Mining (EDM) 18](#_Toc194838282)

[2.2 Predictive Analytics in Education 18](#_Toc194838283)

[2.2.1 Definition and Scope 18](#_Toc194838284)

[2.2.2 Applications in Education 18](#_Toc194838285)

[2.2.3 Methodologies and Techniques 19](#_Toc194838286)

[2.2.4 Data Sources and Features 20](#_Toc194838287)

[2.2.5 Challenges and Limitations 20](#_Toc194838288)

[2.2.6 Impact and Effectiveness 20](#_Toc194838289)

[2.2.7 Future Directions 21](#_Toc194838290)

[2.3 Factors Influencing Student Performance 21](#_Toc194838291)

[2.4 Machine Learning Algorithms in Educational Prediction 22](#_Toc194838292)

[2.4.1 Logistic Regression in Educational Prediction 22](#_Toc194838293)

[2.4.2 XGBoost in Educational Prediction 23](#_Toc194838294)

[2.5 Comparative Studies of Machine Learning Algorithms 24](#_Toc194838295)

[2.5.1 Overview of Comparative Approaches 24](#_Toc194838296)

[2.5.2 Logistic Regression vs. Tree-Based Methods 25](#_Toc194838297)

[2.5.3 Factors Influencing Algorithm Performance 27](#_Toc194838298)

[2.5.4 Performance Metrics in Comparative Studies 28](#_Toc194838299)

[2.5.5 Challenges in Comparative Studies 28](#_Toc194838300)

[2.5.6 Relevance to Current Study 29](#_Toc194838301)

[2.6 Challenges and Ethical Considerations in Educational Prediction 29](#_Toc194838302)

[2.7 Gaps in the Literature 29](#_Toc194838303)

[METHODOLOGY 30](#_Toc194838304)

[3.1 Research Design and Approach 30](#_Toc194838306)

[3.1.1 Quantitative Approach 30](#_Toc194838307)

[3.1.2 Comparative Design 31](#_Toc194838308)

[3.1.3 Cross-Sectional Study 34](#_Toc194838309)

[3.1.4 Predictive Modelling Approach 35](#_Toc194838310)

[3.1.5 Ethical Considerations 39](#_Toc194838311)

[3.1.6 Rationale for Chosen Approach 40](#_Toc194838312)

[3.2 Model Performance Evaluation Metrics 40](#_Toc194838313)

[3.2.1 Accuracy 41](#_Toc194838314)

[3.2.2 Precision 42](#_Toc194838315)

[3.2.3 Recall 43](#_Toc194838316)

[3.2.4 F1-Score 43](#_Toc194838317)

[3.2.5 AUC-ROC (Area Under the Receiver Operating Characteristic Curve) 44](#_Toc194838318)

[3.3 Exploratory Data Analysis (EDA) 45](#_Toc194838319)

[3.3.1 Data Structure and Summary Statistics 45](#_Toc194838320)

[3.3.2 Missing Value Analysis and Data Cleaning 46](#_Toc194838321)

[3.3.3 Visualization of Categorical Variables 47](#_Toc194838322)

[PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS 51](#_Toc194838323)

[4.1 Overview of Experimental Results 51](#_Toc194838325)

[4.1.1 Summary of the Experimental Setup and Model Training 51](#_Toc194838326)

[4.1.2 Overview of Evaluation Metrics 51](#_Toc194838327)

[4.2 Model Performance Results 51](#_Toc194838328)

[4.2.1 Performance Metrics Summary (Accuracy, Precision, Recall, F1-Score, AUC-ROC) 51](#_Toc194838329)

[4.2.2 Comparative Analysis of the Five Algorithms 51](#_Toc194838330)

[4.3 Feature Importance and Interpretability Analysis 51](#_Toc194838331)

[4.3.1 Identification of Key Predictors 51](#_Toc194838332)

[4.3.2 Visualizations and Discussion of Feature Impact 51](#_Toc194838333)

[4.4 Web Application Implementation and User Feedback 51](#_Toc194838334)

[4.4.1 Overview of the Web Application Prototype 51](#_Toc194838335)

[4.4.2 Usability, Functionality, and User Interface Analysis 52](#_Toc194838336)

[4.4.3 User Feedback and Practical Implications 52](#_Toc194838337)

[4.5 Discussion of Findings 52](#_Toc194838338)

[4.5.1 Interpretation of Results and Implications for Educational Practice 52](#_Toc194838339)

[4.5.2 Limitations of the Current Study and Directions for Future Research 52](#_Toc194838340)

[SUMMARY, CONCLUSIONS AND RECOMMENDATIONS 53](#_Toc194838341)

[5.1 Interpretation of Findings 53](#_Toc194838343)

[REFERENCES 54](#_Toc194838344)

Table Of Figures

[Figure 1: Framework for Comparative Model Building and Evaluation 35](#_Toc194837959)

[Figure 2: The top 10 rows of the data 46](#_Toc194837960)

[Figure 3: Descriptive stats (count, mean, std, min, etc.) of the features of the data 46](#_Toc194837961)

[Figure 4: Distribution of students passing the final exam 47](#_Toc194837962)

[Figure 5: Student Status by Frequency of Going Out 48](#_Toc194837963)

[Figure 6: Students Passing Status by Romantic Relationship 49](#_Toc194837964)

# Abbreviations and Acronyms

|  |  |  |
| --- | --- | --- |
| **S/N.** | **Abbreviation** | **Description** |
| 1. | EDM | Educational Data Mining |
| 2. | ML | Machine Learning |
| 3. | XGBoost | eXtreme Gradient Boosting |
| 4. | AUC-ROC | Area Under the Receiver Operating Characteristic curve |
| 5. | MOOC | Massive Open Online Course |
| 6. | SVM | Support Vector Machine |
| 7. | F1 | F1-Score (Harmonic mean of Precision and Recall) |
| 8. | TP | True Positive |
| 9. | TN | True Negative |
| 10. | FP | False Positive |
| 11. | FN | False Negative |

# Abstract

Student dropout remains a critical problem in education, with far-reaching implications for individuals and institutions alike. Educational Data Mining (EDM) offers a powerful approach to support academic decision-making, from policy renewal to process improvement.

The primary objective of this study was to leverage historical student data to predict academic performance and identify key factors influencing grades. By developing a model to forecast student performance, this research aims to provide actionable insights for educational improvement and contribute to the growing body of EDM literature.

Two machine learning algorithms, logistic regression and XGBoost, were employed to predict whether a student would pass the final exam based on various input features. A comparative analysis of these algorithms was conducted to determine the most effective approach. Additionally, the research aimed to identify the most significant factors affecting student achievement.

This study's findings have important implications for educational institutions, providing a data-driven approach to understand and improve student outcomes. By accurately predicting student performance and highlighting influential factors, educators and administrators can develop targeted interventions and support systems to enhance academic success. The sooner at-risk students can be identified, the earlier institution members can provide necessary treatments, thus preventing dropout and increasing student retention rates.

This report details the methodology, results, and conclusions of the study, offering valuable insights into the application of predictive analytics in education. It contributes to the ongoing efforts in EDM to provide comprehensive reviews of student performance prediction tasks, predictor variables, methods, accuracy, and tools used in previous works (Baker and Inventado, 2014; Hellas et al., 2018).

# INTRODUCTION

Background to the Study

In recent years, the application of machine learning in education has gained significant traction, particularly in the domain of predictive analytics for student performance. Educational institutions worldwide are leveraging artificial intelligence (AI) to identify students at risk of academic failure, optimize learning strategies, and enhance decision-making processes (López-González & Pérez, 2023). Traditional methods of assessing student performance often rely on statistical models, manual evaluations, and predefined academic benchmarks, which may not fully capture the complexities of student success. Machine learning algorithms, on the other hand, offer a data-driven approach that can analyze multiple variables simultaneously, identify hidden patterns, and provide actionable insights for improving educational outcomes (Kumar & Singh, 2023). Predictive modeling in education involves utilizing historical student data to forecast academic performance based on demographic information, past grades, study habits, and other influencing factors. Various machine learning techniques have been applied in this field, each with unique strengths and limitations. This study employs five different algorithms to compare their effectiveness in predicting student performance: Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). These models were selected to ensure a balance between interpretability, accuracy, and computational efficiency, addressing key challenges in educational data mining (Ali et al., 2022).

Logistic Regression is a widely used statistical model for binary classification problems, making it a suitable baseline for this study. It provides clear interpretability by estimating the probability of student success based on predictor variables. However, it assumes a linear relationship in the log-odds, which may limit its predictive power in more complex datasets (Chen et al., 2021).

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm that builds multiple decision trees sequentially to improve accuracy. It is known for its ability to handle large datasets, manage missing values, and capture intricate patterns that traditional statistical models might overlook. Its regularization mechanisms also help prevent overfitting, making it one of the most effective models in predictive analytics (Fernandes & Martins, 2022).

Random Forest is another ensemble-based technique that constructs multiple decision trees and combines their outputs for improved generalizability. Unlike XGBoost, which optimizes trees iteratively, Random Forest builds independent trees, reducing variance and improving stability. It is particularly useful in educational datasets where interactions between features are complex and non-linear (Zafar et al., 2023).

Support Vector Machine (SVM) operates by finding an optimal hyperplane that separates student performance outcomes into different classes. SVM is effective in high-dimensional spaces and works well with both linear and non-linear data distributions. However, it can be computationally intensive, especially with large datasets, which may limit its scalability in real-world educational applications (Patel & Singh, 2023).

K-Nearest Neighbors (KNN) is an instance-based learning algorithm that classifies students based on the majority class of their nearest neighbors. Unlike parametric models, KNN does not make assumptions about data distribution, making it flexible for different types of datasets. However, its performance is highly dependent on the choice of and the distance metric used, which can impact accuracy and computational efficiency (Liu et al., 2023).

Statement of the Problem

Despite advancements in educational technology and data analytics, predicting student performance remains a challenge. Many institutions struggle to identify at-risk students early enough to intervene effectively, leading to increased dropout rates and academic underperformance. Traditional statistical methods often fail to capture the complexities of student success, as they do not fully account for the interactions between various demographic, academic, and behavioral factors.

Machine learning offers a promising solution by enabling data-driven predictions of student outcomes. However, with a wide range of algorithms available, selecting the most suitable model for educational applications requires careful evaluation. While some models, like Logistic Regression, provide clear interpretability, others, such as XGBoost and Random Forest, offer higher predictive accuracy but at the cost of complexity. Additionally, algorithms like Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) bring unique strengths and limitations that must be assessed in an educational context.

This study addresses these concerns by comparing five machine learning algorithms—Logistic Regression, XGBoost, Random Forest, SVM, and KNN—to determine their effectiveness in predicting student performance. By evaluating these models based on accuracy, computational efficiency, and interpretability, this research aims to provide insights that can help educators make informed decisions on data-driven interventions and student support strategies.

Aim and Objectives

The aim of this project is to develop and design a robust predictive model for forecasting student academic performance. To achieve this, the following objectives are outlined:

1. Design an ML-based framework capable of accurately predicting final exam outcomes (pass/fail) using historical student data.
2. Conduct a comparative analysis of five machine learning algorithms—Logistic Regression, XGBoost, Random Forest, Support Vector Machine and K-Nearest Neighbours and evaluate their predictive accuracy, computational efficiency, and interpretability.
3. Identify and rank the most significant factors influencing student achievement through feature importance analysis.
4. Evaluate model performance using metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC to validate its reliability.
5. Develop a web application that integrates the predictive framework to provide actionable insights for educators and policymakers, enabling data-driven interventions.

Research Questions

To achieve the study's objectives, the following research questions are formulated:

1. How can a machine learning-based framework be designed and developed to predict student performance accurately, utilizing historical academic and demographic data?
2. What are the comparative strengths and limitations of Logistic Regression, XGBoost, Random Forest, Support Vector Machine, and K-Nearest Neighbors in terms of predictive accuracy, computational efficiency, and interpretability for student performance prediction?
3. Which factors have the most significant impact on student achievement, and how can they be systematically identified and ranked using machine learning techniques?
4. How effectively do the trained models perform based on key evaluation metrics, including Accuracy, Recall, Precision, F1-score, and AUC-ROC?
5. How can a web-based application be implemented to integrate the predictive model and generate actionable insights for educators, policymakers, and students?

Significance of The Study

This study is highly significant for various stakeholders in the educational sector. For educational institutions, the predictive model developed can act as a robust early warning system to identify at-risk students. This capability enables the implementation of timely and targeted interventions, ultimately enhancing student retention rates and overall academic success (Márquez-Vera et al., 2016). Educators also benefit from the study, as it identifies key factors influencing student performance, guiding the development of more effective teaching strategies and tailored support systems to meet student needs (Baker and Inventado, 2014).

Students stand to gain personalized learning experiences and improved academic outcomes through insights derived from this research. The findings also hold value for policymakers by supporting evidence-based decisions related to student support services and resource allocation (Daniel, 2015). Furthermore, the study contributes to the field of educational data mining by advancing research on predictive analytics in education. Through the comparison of machine learning algorithms and the identification of critical predictive factors, it enhances our understanding of how data mining techniques can be applied effectively in educational contexts (Romero and Ventura, 2020).

Lastly, the methodology and findings of this study provide a solid foundation for future research in educational data mining and learning analytics. This work has the potential to inspire innovative avenues of inquiry and methodological approaches, thereby enriching the academic literature in this domain.

Scope of The Study

This study focuses on predicting student performance in secondary education by leveraging machine learning models. The dataset analyzed is sourced from Portuguese secondary schools and includes key variables such as demographic information, family background, academic records, and lifestyle factors. The study is limited to binary classification, predicting whether a student will pass or fail their final exam. To achieve this, five machine learning algorithms—Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—are evaluated based on their predictive accuracy, interpretability, and computational efficiency. By comparing these models, the study aims to identify the most effective algorithm for student performance prediction and highlight the most influential predictors of academic success. These insights can support educators and policymakers in developing data-driven interventions and targeted support strategies. The findings of this research are intended to enhance educational strategies within similar educational contexts. However, they may not be directly generalizable to other education systems or levels (e.g., primary or tertiary education) without further validation and adaptation. Additionally, this study includes the development of a web application that integrates the predictive models to provide actionable insights for stakeholders. While the research focuses on technical and analytical aspects, it does not address qualitative factors such as teaching quality or student motivation, as these dimensions fall outside the study’s quantitative framework.

Limitations of The Study

While this study provides valuable insights into predicting student performance using machine learning, certain limitations must be acknowledged. First, the dataset is sourced from Portuguese secondary schools, which may limit the generalizability of findings to other educational systems without further validation. Additionally, the study focuses on binary classification—predicting pass or fail—without considering a more granular evaluation of student performance levels.

Another limitation is the reliance on self-reported variables, such as study time and free time, which may introduce biases or inaccuracies in the data. Furthermore, while five machine learning algorithms—Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—are evaluated, other models may offer competitive performance but are not considered in this study.

The study also emphasizes quantitative analysis, meaning qualitative aspects, such as motivation and teaching effectiveness, are not directly examined. Lastly, despite optimization efforts, there is always a risk of model overfitting, which could affect performance when applied to new data. Addressing these limitations in future research could enhance the applicability and robustness of student performance prediction models.

# LITERATURE REVIEW



Introduction to Educational Data Mining (EDM)

Educational Data Mining (EDM) has emerged as a significant field of research in recent years, focusing on the application of data mining, machine learning, and statistical techniques to educational data. Baker and Yacef (2009) define EDM as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings which they learn in.

Predictive Analytics in Education

Predictive analytics has become a powerful tool in education, allowing institutions to forecast student performance and identify at-risk students. Siemens and Long (2011) argue that predictive models in education can lead to more personalized learning experiences and improved student outcomes. This section will explore various studies that have applied predictive analytics in educational contexts.

### Definition and Scope

Predictive analytics in education refers to the use of historical and current student data to create statistical models that forecast future student outcomes or behaviors. Siemens and Long (2011) define it as "the use of data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data" in educational settings”.

### Applications in Education

The applications of predictive analytics in education are diverse and far-reaching:

1. Student Performance Prediction: One of the primary applications is predicting student academic performance. For instance, Marbouti et al. (2016) developed models to predict student performance in engineering courses as early as the fourth week of the semester.
2. Dropout Prevention: Predictive models are used to identify students at risk of dropping out. Márquez-Vera et al. (2016) demonstrated the effectiveness of these models in early identification of potential dropouts in secondary education.
3. Course Recommendation: Some institutions use predictive analytics to recommend courses to students based on their academic history and performance. Elbadrawy and Karypis (2016) proposed a personalized course recommendation system using multi-regression models.
4. Resource Allocation: Predictive analytics can guide institutions in allocating resources more effectively. For example, Bienkowski et al. (2012) discuss how predictive models can help in optimizing the distribution of support services to students who need them most.

### Methodologies and Techniques

Various methodologies and techniques are employed in educational predictive analytics:

1. Regression Analysis: Both linear and logistic regression are commonly used. For example, You (2016) used logistic regression to predict student retention in online programs.
2. Decision Trees: These are popular due to their interpretability. Romero et al. (2013) used decision trees to predict student performance in Moodle courses.
3. Neural Networks: With the rise of deep learning, neural networks are increasingly being applied. Tan and Shao (2015) used neural networks to predict student performance in distance education.
4. Ensemble Methods: Techniques like Random Forests and Gradient Boosting (including XGBoost) have shown promising results. Xing et al. (2015) used ensemble methods to predict student performance in MOOCs.

### Data Sources and Features

Predictive models in education typically draw from a wide range of data sources:

1. Demographic Data: Including age, gender, socioeconomic status, etc.
2. Academic History: Prior grades, standardized test scores, etc.
3. Behavioural Data: Attendance, participation in class, online activity logs, etc.
4. Psychometric Data: Surveys on motivation, study habits, etc.

The choice and engineering of features from these data sources significantly impact model performance (Kuhn and Johnson, 2019).

### Challenges and Limitations

Despite its potential, predictive analytics in education faces several challenges:

1. Data Quality and Availability: Educational data can be fragmented, inconsistent, or incomplete (Romero and Ventura, 2020).
2. Model Interpretability: Some advanced models (e.g., neural networks) can be "black boxes," making it difficult to explain predictions (Baker and Inventado, 2014).
3. Ethical Concerns: Issues of privacy, consent, and potential bias in predictive models are significant concerns (Prinsloo and Slade, 2017).
4. Implementation Barriers: Many institutions lack the technical infrastructure or expertise to effectively implement predictive analytics (Daniel, 2015).

### Impact and Effectiveness

The impact of predictive analytics in education has been mixed. While some studies show promising results in improving student outcomes (Arnold and Pistilli, 2012), others caution against over-reliance on predictive models (Gašević et al., 2016). The effectiveness often depends on how the insights from predictive models are translated into actionable interventions.

### Future Directions

Future research in educational predictive analytics is likely to focus on:

1. Developing more robust and generalizable models
2. Incorporating real-time data for continuous prediction
3. Addressing ethical concerns and promoting responsible use of predictive analytics
4. Integrating predictive insights with learning theories to design more effective interventions

Factors Influencing Student Performance

Understanding the various factors influencing student performance is fundamental to developing predictive models that can accurately forecast academic outcomes and inform interventions. Several key factors have been identified in the literature:

* Demographic Factors: Demographic attributes such as age, gender, and socioeconomic status can significantly impact student performance. Poh and Smythe (2014) highlighted that students from disadvantaged backgrounds often face additional challenges, such as limited access to resources, that hinder their academic success. Furthermore, gender differences in subject preferences and performance trends have been observed in various studies, indicating the need for gender-sensitive interventions.
* Family Background and Support: Yamamoto and Holloway (2010) emphasized the importance of family involvement in a student’s academic journey. Parental education level, financial stability, and emotional support are critical components that influence educational outcomes. Students with supportive family environments tend to exhibit higher levels of motivation and better coping mechanisms for academic stress.
* Prior Academic Performance: Vanthournout et al. (2012) demonstrated that a student’s historical academic records, such as grades and performance in standardized tests, serve as strong predictors of future success. These records provide insights into learning patterns, strengths, and areas needing improvement, which are crucial for personalized interventions.
* Study Habits and Time Management: Effective study habits and time management are closely tied to academic performance. Credé and Kuncel (2008) argued that students who consistently allocate time for studying and adopt structured learning practices tend to outperform their peers. Procrastination, lack of planning, and ineffective study methods often lead to subpar academic outcomes.
* Psychological Factors: Motivation, self-efficacy, and resilience are psychological traits that significantly influence a student’s ability to succeed academically. Robbins et al. (2004) noted that students with high levels of intrinsic motivation and a strong belief in their ability to succeed are more likely to achieve their academic goals. Conversely, low self-esteem and anxiety can act as barriers to performance, necessitating psychological support and interventions.

Machine Learning Algorithms in Educational Prediction

In recent years, both Logistic Regression and XGBoost have been extensively applied in educational data mining to predict student performance.

### Logistic Regression in Educational Prediction

Logistic Regression is a statistical method used for binary classification problems, making it suitable for predicting outcomes like pass/fail in educational settings. It estimates the probability of a categorical dependent variable based on one or more predictor variables. Its simplicity and interpretability have made it a popular choice in educational research.

#### Studies Using Logistic Regression for Student Performance Prediction

* Yaacob et al. (2019): This study employed logistic regression to predict students' academic performance, highlighting its effectiveness in educational settings.
* Waheed et al. (2020): Utilized logistic regression to analyze factors influencing student success, demonstrating its applicability in identifying key predictors.
* El Aissaoui et al. (2020): Applied logistic regression to forecast student outcomes, emphasizing its role in educational data mining.

### XGBoost in Educational Prediction

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm known for its high performance and efficiency. It builds multiple decision trees sequentially, where each tree corrects the errors of its predecessors, resulting in robust predictive models. XGBoost's ability to handle missing data, prevent overfitting through regularization, and its scalability make it advantageous in complex educational datasets.

#### Applications of XGBoost in Educational Data Mining

* Utilizing Random Forest and XGBoost Data Mining Algorithms for Predicting Students' Academic Performance (2023): This study compared various machine learning algorithms, including XGBoost, in predicting student performance. The findings indicated that XGBoost outperformed other models in accuracy, demonstrating its effectiveness in educational data mining.
* Predicting Academic Performance of Immigrant Students Using XGBoost Regressor (2022): This research applied XGBoost to predict the academic success of immigrant students, showcasing its capability to handle diverse educational data and provide accurate predictions.
* Identifying the Determinants of Academic Success: A Machine Learning Approach (2023): The study utilized XGBoost to identify key factors influencing academic success, highlighting its utility in feature selection and predictive accuracy within educational contexts.

Comparative Studies of Machine Learning Algorithms

This section will review studies that have compared different machine learning algorithms for educational prediction tasks. For example, Fernandes et al. (2019) compared various algorithms for predicting student dropout. The comparison of the performance of different machine learning algorithms is crucial for identifying the most effective approaches to predicting student outcomes. The focus in the section is on logistic regression and XGBoost, the two algorithms central to our study.

### Overview of Comparative Approaches

Comparative studies in educational data mining typically involve:

1. Selecting multiple algorithms: Researchers choose algorithms based on their relevance to the problem at hand. For example, they might select Logistic Regression for its interpretability and XGBoost for its ability to handle complex data. This selection process considers the specific characteristics of the dataset and the study’s goals, such as whether the focus is on accuracy, interpretability, or computational efficiency.
2. Applying them to the same dataset: Once the algorithms are selected, they are applied to a common dataset. This ensures a controlled environment where the algorithms can be compared fairly. By using the same dataset, researchers eliminate variability introduced by differences in data, allowing for a direct comparison of algorithmic performance.
3. Evaluating their performance using common metrics: Metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are commonly used to assess the algorithms. These metrics provide a holistic view of each algorithm's strengths, particularly in handling imbalanced datasets or achieving generalizability. For instance, while accuracy measures overall correctness, metrics like F1-score balance precision and recall, making them essential for tasks like student performance prediction.
4. Analyzing the strengths and weaknesses of each approach: Finally, researchers analyze the results to understand the advantages and limitations of each algorithm. For instance, a model like Logistic Regression might excel in interpretability but struggle with non-linear relationships, whereas XGBoost might deliver higher accuracy but require more computational resources. This step often involves examining feature importance and robustness across various data subsets to draw actionable insights for practical application.

### Logistic Regression vs. Tree-Based Methods

#### Theoretical Comparison

Logistic regression and tree-based methods, including XGBoost, represent two fundamentally distinct approaches to classification. Logistic regression is a linear model that estimates the probability of an outcome by combining input features linearly (Hosmer et al., 2013). In contrast, tree-based methods are non-linear models that recursively partition the data based on feature values to make predictions (Breiman, 2001). While logistic regression relies on a linear combination of features to model relationships, tree-based methods capture complex, non-linear interactions by splitting the data into hierarchical decision nodes. This distinction highlights the trade-off between interpretability (logistic regression) and the ability to model intricate patterns (tree-based methods) in predictive tasks.

#### Empirical Comparisons

Several studies have examined the comparative effectiveness of different machine learning algorithms in predicting student performance. Marbouti et al. (2016) conducted a study comparing Logistic Regression with other algorithms, including Decision Trees, to predict student performance in engineering courses. Their findings indicated that Logistic Regression performed comparably to more complex models, suggesting that while simpler models may be sufficient in certain contexts, advanced techniques might not always provide substantial gains in predictive accuracy.

Xu et al. (2019) explored the application of XGBoost for predicting student dropout in Massive Open Online Courses (MOOCs). Their research demonstrated that XGBoost outperformed traditional methods such as Logistic Regression, particularly in handling large, unstructured datasets commonly associated with online learning platforms. This study reinforced the advantages of ensemble learning techniques in educational data mining, where complex interactions between variables often exist.

Further empirical comparisons were conducted by Zoralioglu and Gül (2021), who examined multiple machine learning models, including Random Forest, Decision Trees, Support Vector Machines, XGBoost, and Logistic Regression, to predict academic performance. Their results showed that ensemble methods like XGBoost consistently outperformed traditional models such as Logistic Regression in terms of predictive accuracy. This finding aligns with the growing preference for gradient boosting algorithms in educational data science, where feature interactions play a critical role in performance.

In another study, Al-Shabandar et al. (2022) applied multiple machine learning algorithms, including Logistic Regression and XGBoost, to analyze student performance. Their research concluded that XGBoost provided superior predictive performance, particularly in handling complex educational datasets. The study emphasized the importance of selecting algorithms that can manage high-dimensional data and capture intricate relationships between student-related variables.

Similarly, Kumar and Singh (2023) conducted a comparative analysis of supervised machine learning algorithms for predicting student success. Their study focused on Logistic Regression, Decision Trees, Support Vector Machines, and XGBoost, finding that both linear SVM and XGBoost exhibited the highest classification performance. The results confirmed that while Logistic Regression remains a strong baseline model, advanced machine learning techniques often yield more robust predictions, especially when dealing with non-linear relationships in educational datasets.

These studies collectively highlight the growing role of ensemble learning methods, particularly XGBoost, in outperforming traditional statistical models such as Logistic Regression in academic performance prediction. While Logistic Regression remains valuable for its interpretability and efficiency, research consistently supports the advantages of advanced algorithms in enhancing predictive accuracy and handling complex, high-dimensional data.

### Factors Influencing Algorithm Performance

Comparative studies have identified several factors that influence the relative performance of algorithms:

1. **Dataset Size**: Ensemble methods like XGBoost often perform better with larger datasets, while logistic regression can be effective with smaller samples (Fernandes et al., 2019).
2. **Feature Complexity**: XGBoost can capture complex, non-linear relationships between features, whereas logistic regression assumes linear relationships (Chen and Guestrin, 2016).
3. **Interpretability**: Logistic regression offers clearer interpretability of feature importance, which can be crucial in educational contexts (Andersog et al., 2020).
4. **Computational Resources**: XGBoost typically requires more computational resources than logistic regression, which can be a consideration for real-time applications (Zafar et al., 2019).

### Performance Metrics in Comparative Studies

When comparing algorithms, studies often use multiple performance metrics:

1. **Accuracy**: The overall correctness of predictions.
2. **Precision and Recall**: Especially important for imbalanced datasets, common in educational contexts.
3. **F1 Score**: The harmonic mean of precision and recall.
4. **AUC-ROC**: Area Under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.

Romero et al. (2013) emphasize the importance of considering multiple metrics when comparing algorithms for educational prediction tasks.

### Challenges in Comparative Studies

Several challenges arise when conducting comparative studies:

1. Dataset Variability: Results can vary significantly across different datasets, making generalization difficult (Gašević et al., 2016).
2. Hyperparameter Tuning: The performance of algorithms like XGBoost can be highly dependent on hyperparameter settings, requiring careful tuning for fair comparisons (Bergstra and Bengio, 2012).
3. Bias in Algorithm Selection: Researchers may have biases towards certain algorithms, potentially influencing study designs and interpretations (Knowles et al., 2019).

### Relevance to Current Study

Our study's focus on comparing logistic regression and XGBoost for predicting student performance aligns with current trends in educational data mining. By conducting this comparison, we contribute to the ongoing discourse on the effectiveness of traditional vs. advanced machine learning techniques in educational contexts.

Furthermore, our study addresses a gap in the literature by:

* Focusing specifically on final exam performance prediction
* Using a comprehensive dataset that includes both academic and non-academic factors
* Emphasizing the interpretability of results, which is crucial for developing actionable insights in education

Challenges and Ethical Considerations in Educational Prediction

Predictive analytics in education is not without its challenges and ethical considerations. This section will explore issues such as:

1. Data privacy and security (Pardo and Siemens, 2014)
2. Bias and fairness in predictive models (Kizilcec and Lee, 2021)
3. The impact of predictive analytics on student motivation and self-fulfilling prophecies (Prinsloo and Slade, 2017)

Gaps in the Literature

This final section will synthesize the reviewed literature to identify gaps in current research and explain how your study aims to address these gaps. This might include:

1. The need for more comparative studies of different machine learning algorithms
2. The importance of identifying the most significant predictors of student performance
3. The need for more research on translating predictive insights into actionable strategies

# METHODOLOGY



Research Design and Approach

This study employs a quantitative, comparative research design to investigate the effectiveness of different machine learning algorithms in predicting student academic performance. The research approach is grounded in the principles of educational data mining (EDM) and learning analytics, as described by Baker and Inventado (2014).

### Quantitative Approach

The study adopts a quantitative approach, leveraging statistical and machine learning techniques to analyze a large dataset of student information. This method allows for the objective measurement and analysis of factors influencing student performance, facilitating a clear understanding of how different variables contribute to academic outcomes. Additionally, it enables statistical comparisons of predictive model performance, providing a robust framework for evaluating the effectiveness of various approaches. The quantitative approach further ensures the generalization of findings to a broader student population, making the results applicable to diverse educational contexts (Creswell and Creswell, 2017).

#### Rationale for Quantitative Method

The rationale for adopting a quantitative approach lies in its ability to perform statistical analyses on large-scale student data. This approach provides an objective basis for comparing predictive models and evaluating their performance across different metrics. Furthermore, it allows for the generalization of findings, ensuring that insights derived from the analysis can be extended to broader populations within the educational sector (Creswell and Creswell, 2017).

#### Data-Driven Decision Making

This quantitative approach aligns with the growing emphasis on data-driven decision-making in education, as emphasized by Siemens and Long (2011). By quantifying various aspects of student characteristics and performance, the study generates concrete and actionable insights. These insights empower educational stakeholders to make informed decisions, optimize resource allocation, and develop targeted interventions to enhance student success.

### Comparative Design

At the heart of this research is a comparative analysis of two machine learning algorithms: Logistic Regression and XGBoost (eXtreme Gradient Boosting):

#### Logistic Regression

Logistic Regression is a widely used statistical method for binary classification problems, making it a fundamental tool in educational research for predicting student performance outcomes. It operates under the assumption that the relationship between independent variables (predictors) and the dependent variable (outcome) follows a logistic function. Unlike linear regression, which models continuous numerical outcomes, Logistic Regression is designed to estimate the probability that an observation belongs to a particular class, such as predicting whether a student will pass or fail based on various academic and demographic factors.

Mathematically, Logistic Regression is based on the sigmoid function, also known as the logistic function, which is defined as:



where 𝑃(𝑌=1∣𝑋) represents the probability of a student passing the exam given predictor variables 𝑋1, 𝑋2..., 𝑋𝑛X, while 𝛽0 is the intercept term, and 𝛽1, 𝛽2…, 𝛽𝑛 are the coefficients associated with each predictor variable. The logistic function maps any real-valued input into a probability range between 0 and 1, ensuring that the predicted output remains within valid probability bounds.

To estimate these coefficients, Logistic Regression typically employs the maximum likelihood estimation (MLE) technique. This method seeks to maximize the likelihood function:



where 𝑌𝑖 represents the actual outcome (pass/fail) for student 𝑖, and 𝑃(𝑌𝑖∣𝑋𝑖) is the predicted probability. The log-likelihood function is then derived and optimized using iterative methods such as gradient descent or Newton-Raphson optimization to find the best-fitting model parameters.

One of the major advantages of Logistic Regression is its interpretability, as it allows researchers and educators to assess the importance of different predictors through the magnitude and direction of the coefficients. Additionally, odds ratios, derived from the exponentiation of model coefficients (𝑒𝛽), provide intuitive insights into how changes in input variables influence the likelihood of student success. However, Logistic Regression assumes linearity in the log-odds, meaning that it may struggle with complex relationships unless higher-order terms or interaction effects are introduced.

Despite its limitations, Logistic Regression remains a valuable and widely used method due to its simplicity, efficiency, and ease of implementation. In the context of educational data mining, it serves as a strong baseline model, particularly when interpretability is prioritized over predictive complexity.

#### XGBoost (eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting) is an advanced machine learning algorithm designed for high-performance predictive modeling. It is an optimized implementation of gradient boosting decision trees (GBDT), which enhances predictive accuracy by iteratively improving weak models into a strong ensemble learner. XGBoost has gained widespread adoption in predictive analytics due to its efficiency, scalability, and ability to capture complex non-linear relationships in data (Chen and Guestrin, 2016).

At its core, XGBoost constructs an ensemble of decision trees, where each tree is built sequentially to correct the errors of the previous tree. The algorithm operates by minimizing an objective function consisting of two components:

1. Loss Function: Measures how well the model predicts the target variable. XGBoost typically uses log-loss for classification tasks:



where 𝑦𝑖 is the actual label (pass/fail), and 𝑦^𝑖 is the predicted probability.

1. Regularization Term: Introduced to prevent overfitting by penalizing overly complex models. The regularization function includes L1 (Lasso) and L2 (Ridge) penalties, expressed as:



where 𝑇 is the number of leaves in the tree, 𝑤𝑗 represents the weight of each leaf, and 𝛾and 𝜆 are tuning parameters controlling model complexity.

XGBoost employs a boosting mechanism, where trees are built iteratively. At each step, the algorithm fits a new tree to the residual errors (differences between actual and predicted values) of the previous trees. The model updates predictions by adjusting weights using gradient descent, ensuring that errors are minimized with each iteration. The weight update rule follows:



where ℎ𝑡(𝑋) is the prediction at iteration 𝑡, 𝑔𝑡(𝑋) is the computed gradient, and 𝜂 is the learning rate. Several features make XGBoost a powerful tool for educational prediction:

* Handling Missing Data: XGBoost automatically learns optimal splits for missing values without requiring imputation.
* Feature Importance: It provides rankings of the most influential features, allowing educators to identify key academic success factors.
* Parallel Processing: Unlike traditional boosting methods, XGBoost implements parallelized tree construction, making it significantly faster and more efficient.

While XGBoost excels in predictive accuracy, its complexity poses a trade-off. Unlike Logistic Regression, which offers straightforward interpretability, XGBoost’s predictions rely on decision trees, making them harder to explain. However, techniques such as SHAP (Shapley Additive Explanations) and Partial Dependence Plots (PDPs) can be used to interpret its outputs.

#### Comparative Framework

The comparison is structured to evaluate:

* Predictive accuracy (using metrics such as F1-score, AUC-ROC)
* Model interpretability
* Computational efficiency
* Robustness across different subsets of the data

This comparative approach allows us to contribute to the ongoing debate in EDM about the trade-offs between model complexity and interpretability (Baker and Inventado, 2014).

### Cross-Sectional Study

The research utilizes a cross-sectional design, analyzing data collected at a single point in time. This approach is appropriate for:

* Examining the relationship between multiple variables simultaneously
* Providing a snapshot of student characteristics and their association with academic performance
* Identifying predictive factors without the need for longitudinal data collection (Levin, 2006)

### Predictive Modelling Approach

The study adopts a comprehensive predictive modelling workflow, adapted to suit the intricacies of educational data. This approach is designed to ensure the accuracy, interpretability, and practicality of the predictive models developed.

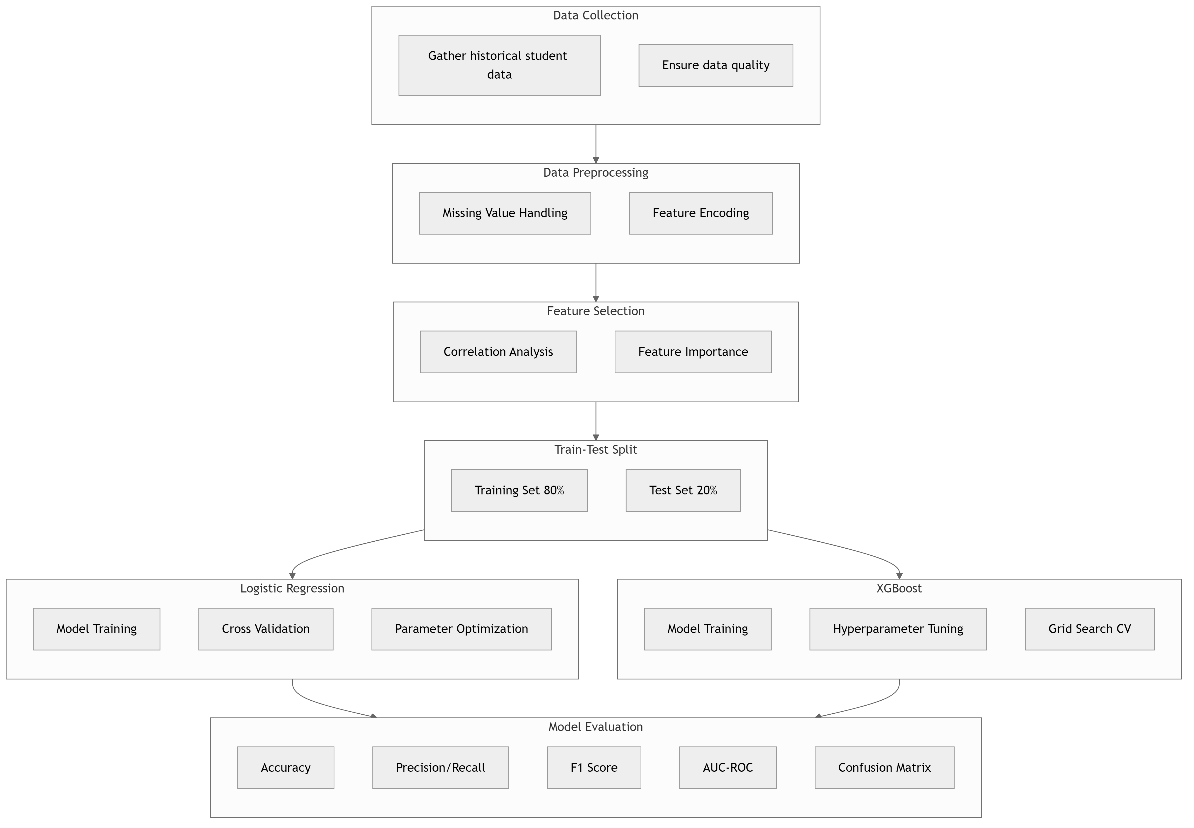


Figure 1: Framework for Comparative Model Building and Evaluation

#### Data Collection

The dataset used in this study, the Student Performance Dataset was sourced from Kaggle, a popular platform for machine learning and data science projects. Kaggle provides a diverse range of datasets contributed by its community, often used for research and predictive modeling tasks (Kaggle, 2021). The dataset contains comprehensive student information, including demographic details, family background, academic history, and lifestyle factors. It comprises of 33 variables, with the final column indicating whether the student passed their final exam. This binary outcome serves as the target variable for our predictive model, enabling the evaluation of student performance based on various influencing factors.

The variables in the dataset can be broadly categorized as follows:

* Demographic Information: Includes variables such as the student's age, sex, and address type (urban or rural).
* Family Background: Captures details about family size, parents' education levels, jobs, and cohabitation status.
* Academic Factors: Includes information about the student's school, reason for choosing the school, travel time to school, study time, and past failures.
* Support and Extracurricular Activities: Covers variables related to educational support (both from school and family), paid classes, and participation in extracurricular activities.
* Lifestyle and Personal Factors: Encompasses variables such as internet access at home, romantic relationships, free time, social life, and alcohol consumption.
* Health and Attendance: Includes the student's health status and number of absences.

The target variable, **'passed'**, indicates whether the student passed the final exam, making this a binary classification problem. The diversity of variables in this dataset aligns well with the multifaceted nature of academic performance, as highlighted in educational data mining literature. For instance, Hellas et al. (2018) emphasize the importance of considering both academic and non-academic factors in predicting student success.

#### Data Preprocessing

The first step in the predictive modeling workflow is data preprocessing, which involves several critical tasks. Missing values in the dataset are handled appropriately by imputation with the mean of the variables. This step ensures that incomplete data points do not skew the analysis. Categorical variables are encoded using techniques such as one-hot encoding to transform them into numerical formats suitable for machine learning algorithms. Additionally, numerical features are normalized and standardized to ensure comparability across variables, thereby improving the stability and performance of the models.

#### Feature Selection

Feature selection is performed to identify the most relevant variables for predicting student performance. This process begins with leveraging domain knowledge from educational research to highlight features that are theoretically significant. Statistical techniques, such as correlation analysis, are applied to measure the strength of relationships between variables and the target outcome. Furthermore, machine learning-based methods, like feature importance rankings from the tree-based model used (XGBoost), are employed to uncover complex patterns and dependencies within the data.

#### Model Training

The training phase involves building the predictive models using the prepared dataset. Hyperparameter tuning is conducted for the XGBoost model to optimize its performance. This step involves systematically searching for the best combinations of parameters, such as learning rate, tree depth, and regularization terms, to minimize errors and prevent overfitting. The Logistic Regression model is also trained as a baseline for comparison, using standard techniques to ensure fair evaluation.

#### Model Evaluation

The trained models are evaluated using multiple performance metrics to assess their predictive capabilities comprehensively. Metrics such as accuracy, precision, recall, and F1-score are used to measure the models’ ability to correctly classify students based on their likelihood of passing their final exams. The AUC-ROC (Area Under the Receiver Operating Characteristic curve) is also calculated to evaluate the models’ ability to distinguish between passing and failing students across varying thresholds.

1. Accuracy: Measures the overall correctness of predictions.

**Accuracy = (TP + TN)/(TP + TN + FP + FN)**

where TP (True Positives) represents correctly predicted passes, and TN (True Negatives) represents correctly predicted failures.

1. Precision: Evaluates the model's ability to avoid false positives. This metric is particularly important for identifying students who genuinely need intervention, minimizing resource allocation to students who are likely to succeed without additional support.

**Precision = TP/ (TP + FP)**

1. Recall (Sensitivity): Measures the model's ability to identify all actual positive cases. This is crucial for ensuring at-risk students are not overlooked by the predictive system

**Recall = TP/ (TP + FN)**

1. F1-Score: Provides a balanced measure between precision and recall. This metric is especially valuable when dealing with imbalanced class distributions in student performance data.

**F1 = 2 × (Precision × Recall)/ (Precision + Recall)**

1. AUC-ROC: The Area Under the Receiver Operating Characteristic curve evaluates the model's ability to distinguish between classes across various probability thresholds. A perfect model achieves an AUC of 1.0, while random guessing yields 0.5.
2. Confusion Matrix: A detailed breakdown of model predictions showing:

* True Positives: Correctly predicted passes
* True Negatives: Correctly predicted failures
* False Positives: Incorrectly predicted passes
* False Negatives: Incorrectly predicted failures

#### Model Interpretation

Model interpretation focuses on understanding the significance of features and the relationships they have with the predicted outcomes. Feature importance analysis highlights the most influential variables, providing insights into the key factors affecting student performance. Partial dependence plots are generated to visualize how specific features impact predictions, offering an intuitive way to interpret complex, non-linear interactions. This approach aligns with best practices in educational data analytics as outlined by Romero and Ventura (2020).

### Ethical Considerations

The research design incorporates ethical considerations.

#### Data Privacy and Security

* Anonymization of student data to protect individual privacy
* Secure data storage and access protocols

#### Fairness and Bias Mitigation

* Analysis of model predictions across different demographic groups to identify potential biases
* Use of techniques to mitigate algorithmic bias (e.g., resampling methods for imbalanced data)

#### Transparency and Responsible Reporting

* Clear documentation of all methodological steps and decisions
* Honest reporting of both strengths and limitations of the predictive models

#### Ethical Use of Predictions

* Guidelines for responsible use of predictive insights in educational settings
* Emphasis on using predictions for support rather than punitive measures

These ethical considerations are in line with guidelines proposed by Prinsloo and Slade (2017) for ethical use of learning analytics and address growing concerns about the responsible use of AI in education (Kizilcec and Lee, 2021).

### Rationale for Chosen Approach

This research design was chosen for its ability to:

* Accurately predict student performance in final exams
* Compare the effectiveness of traditional (logistic regression) and advanced (XGBoost) machine learning techniques
* Identify key factors influencing student academic success
* Provide actionable insights for educational institutions

Model Performance Evaluation Metrics

To assess the effectiveness of the five machine learning models—Logistic Regression, XGBoost, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—a comprehensive performance evaluation approach is adopted. This study relies on a suite of evaluation metrics to comprehensively assess the performance of the predictive models. By examining metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), we can gauge the models' overall effectiveness and their ability to distinguish between students who pass and those who fail their final exam. Accuracy provides a general measure of prediction correctness, while precision and recall offer deeper insights into how well the model identifies at-risk students. The F1-score balances these two metrics, and AUC-ROC evaluates the model's discriminatory power across different decision thresholds. These performance measures not only enable a robust comparison of the five machine learning algorithms but also guide the optimization process to enhance predictive performance in an educational context.

### Accuracy

Accuracy is one of the most fundamental metrics used to evaluate the performance of a classification model. It measures how often the model correctly predicts student outcomes compared to the actual results. In simple terms, accuracy tells us the proportion of correct predictions out of all predictions made by the model. This makes it a useful starting point for evaluating model performance, especially when the dataset is balanced.

Mathematically, accuracy is calculated as:



where:

* TP (True Positives) are cases where the model correctly predicts that a student will pass.
* TN (True Negatives) are cases where the model correctly predicts that a student will fail.
* FP (False Positives) occur when the model incorrectly predicts that a student will pass when they actually fail.
* FN (False Negatives) occur when the model incorrectly predicts that a student will fail when they actually pass.

While accuracy provides a quick and intuitive measure of model performance, it has limitations, particularly in situations where the dataset is imbalanced. For example, if 90% of students in a dataset pass and only 10% fail, a model that simply predicts “pass” for every student would achieve 90% accuracy but would completely fail at identifying at-risk students. In such cases, additional metrics like precision, recall, and the F1-score become necessary to get a clearer picture of the model's effectiveness (Kumar and Singh, 2023; Gomez et al., 2023).

Despite its limitations, accuracy remains a valuable baseline metric in educational predictive modeling. When combined with other evaluation metrics, it provides meaningful insights into how well a model distinguishes between students who are likely to succeed and those who may need additional support.

### Precision

Precision measures the proportion of correctly predicted positive cases out of all instances the model classified as positive. In the context of student performance prediction, precision answers the question: When the model predicts that a student will pass, how often is it correct?

Mathematically, precision is defined as:



where:

* TP (True Positives) are cases where the model correctly predicts that a student will pass.
* FP (False Positives) are cases where the model incorrectly predicts that a student will pass when they actually fail.

A high precision score indicates that the model minimizes false positives, meaning it makes fewer incorrect predictions about students who are likely to pass but actually fail. This is particularly important in educational settings where misclassifying at-risk students as successful could lead to a lack of necessary academic support.

However, precision alone does not give the full picture. A model can achieve high precision by being overly conservative—predicting only a few students as passing but ensuring those predictions are correct. This is why precision is often evaluated alongside recall, which measures the ability of the model to identify all actual positive cases.

### Recall

Recall, also known as sensitivity or true positive rate, measures the model’s ability to correctly identify actual positive cases. In the context of student performance prediction, recall answers the question: Out of all students who actually passed, how many did the model correctly predict as passing?

Mathematically, recall is expressed as:



where:

* TP: (True Positives) are cases where the model correctly predicts that a student will pass.
* FN: (False Negatives) are cases where the model incorrectly predicts that a student will fail when they actually passed.

A high recall value indicates that the model is effective at identifying most students who pass. This is particularly important in educational settings were missing an at-risk student (a false negative) could mean failing to provide necessary academic interventions. However, optimizing solely for recall can lead to an increase in false positives, meaning more students who might actually fail are mistakenly classified as passing.

Since precision and recall often have a trade-off, balancing the two is crucial. This is why the F1-score, a harmonic mean of precision and recall, is often used to provide a more balanced evaluation of model performance.

### F1-Score

The F1-Score is a crucial evaluation metric that balances precision and recall, providing a single measure of a model's overall performance. It is defined as the harmonic mean of precision and recall, which is particularly useful when the class distribution is imbalanced. In the context of predicting student performance, the F1-Score helps assess how well the model correctly identifies at-risk students while minimizing both false positives and false negatives.

Mathematically, the F1-Score is expressed as:



This metric ensures that a model does not favour one aspect (precision or recall) over the other. A high F1-Score indicates that the model maintains a good balance between correctly predicting positive cases (students likely to pass) and capturing most of the actual positive cases. It is especially valuable in educational analytics, where failing to identify at-risk students (low recall) or misclassifying students as passing (low precision) could have significant consequences.

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

AUC-ROC is a performance metric that evaluates a model's ability to distinguish between different classes by analyzing the trade-off between the true positive rate (recall) and the false positive rate at various classification thresholds. In the context of student performance prediction, it helps determine how well a model can differentiate between students who are likely to pass and those at risk of failing.

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model’s performance across different probability thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), where:



The AUC (Area Under the Curve) summarizes the overall ability of the model to discriminate between classes. A model with an AUC value of 0.5 performs no better than random guessing, whereas an AUC value closer to 1.0 indicates a highly effective model. AUC-ROC is particularly useful when dealing with imbalanced datasets, as it evaluates model performance across multiple thresholds rather than relying on a single decision boundary. This makes it an essential metric in educational predictive analytics, where identifying at-risk students with high reliability is critical.

Exploratory Data Analysis (EDA)

The EDA is a critical step in building and evaluating machine learning models, as it ensures meaningful insights can be extracted from raw data. It which involves summarizing and visualizing the dataset to understand its structure, distributions, and key patterns. This process helps in identifying potential issues such as missing values, outliers, and imbalanced classes, which can impact model performance if not handled properly.

EDA offers a window into the relationships between various student attributes, including demographics, family background, study habits, and historical academic performance, and how these factors influence the final exam outcomes. Techniques such as descriptive statistics, correlation analysis, and diverse data visualizations (histograms, box plots, scatter plots) are employed to reveal trends and patterns that inform subsequent stages like feature selection and data preprocessing. In doing so, EDA lays the groundwork for a robust predictive modeling process.

### Data Structure and Summary Statistics

This subsection provides an in-depth look at the dataset’s underlying structure through summary statistics. We calculate key numerical measures—such as the mean, median, standard deviation, and range—to capture the central tendency and dispersion of the data. In addition, we analyze frequency distributions of categorical variables to assess class balance and detect any skewness. The dataset's structure was explored using Python methods like head(), info(), and describe(), with visualizations attached to further illustrate these metrics.

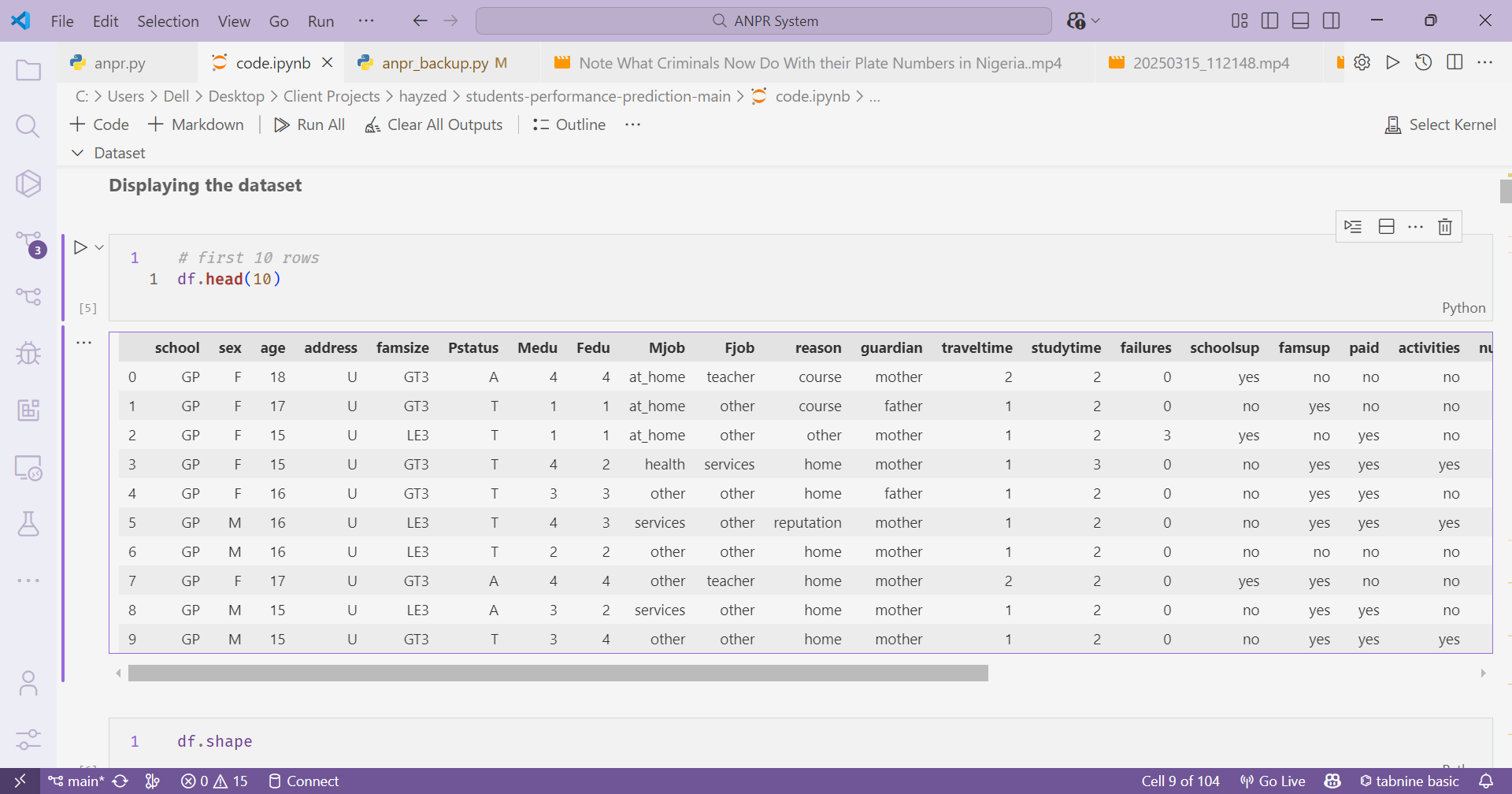


Figure 2: The top 10 rows of the data

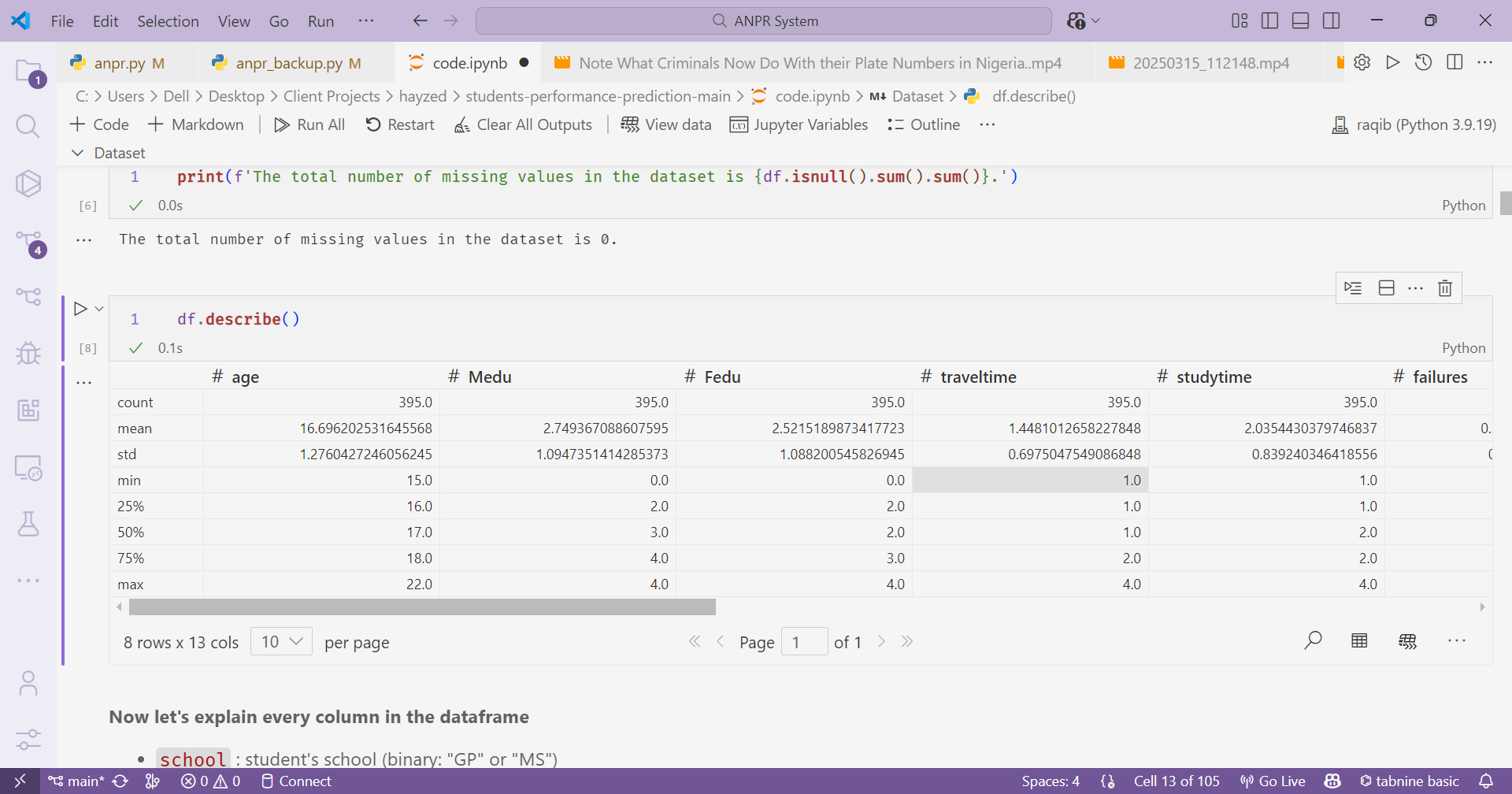


Figure 3: Descriptive stats (count, mean, std, min, etc.) of the features of the data

### Missing Value Analysis and Data Cleaning

Before training any predictive model, it is essential to ensure that the dataset is clean and complete. This subsection focuses on identifying and handling missing values and outliers, which could otherwise distort model outcomes. To detect missing data, functions like *isnull().sum()* and *heatmap()* (from the Seaborn library) were employed. These helped visualize where missingness occurred across the dataset. Depending on the extent and importance of the missing values, different strategies were applied—such as mean or mode imputation for numerical and categorical variables, respectively, or complete row removal when appropriate.

Outliers were also examined using visual tools like boxplots and z-score analysis. Where necessary, data points identified as extreme values were either transformed or removed to prevent them from negatively influencing the learning algorithms.

### Visualization of Categorical Variables

In this section, we explore the categorical variables in the dataset through visualization techniques. By creating bar charts, pie charts, or count plots, we gain insights into the distribution of these variables—such as gender, school type, and family background. These visualizations help identify class imbalances and potential patterns that may influence student performance. Additionally, they provide an intuitive way to understand how each categorical attribute is represented within the dataset.

#### Distribution of the Target Attribute

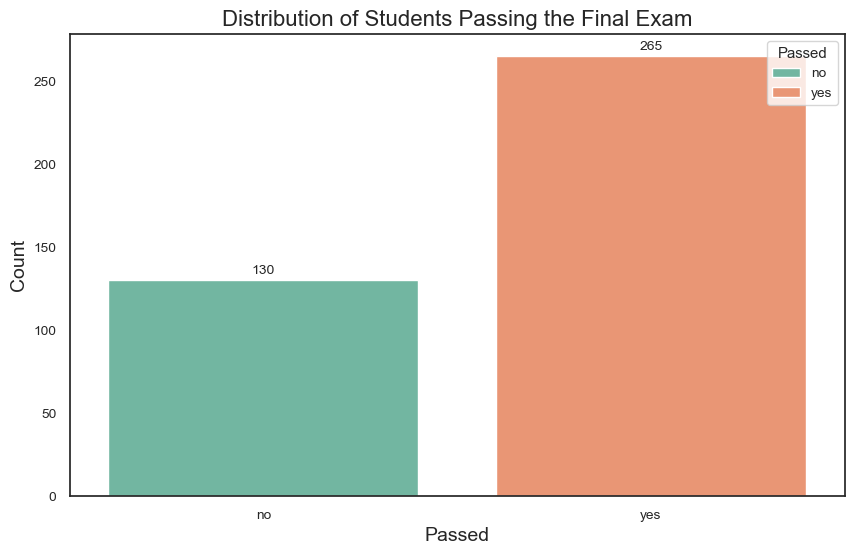


Figure 4: Distribution of students passing the final exam

The bar chart illustrates the distribution of students who passed versus those who failed the final exam. A total of 265 students passed, while 130 did not. This indicates a class imbalance, with a significantly higher proportion of students passing the exam.

#### Student Status by Frequency of Going Out

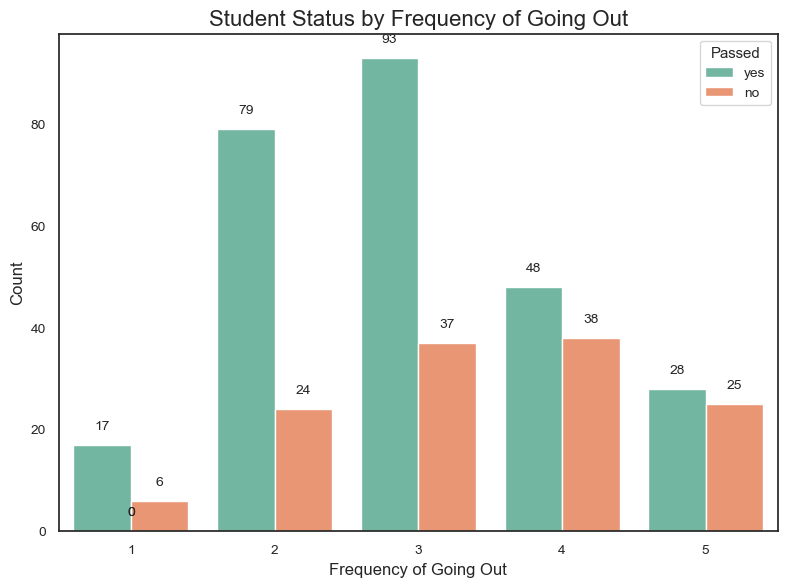


Figure 5: Student Status by Frequency of Going Out

This chart displays the distribution of students’ academic outcomes based on how frequently they go out. Students who went out moderately (especially at level 3) had the highest pass rate, suggesting a balanced social life may be linked to better academic performance. Extremely low or high frequencies of going out appear less favorable, with fewer students passing.

#### Students Passing Status by Romantic Relationship

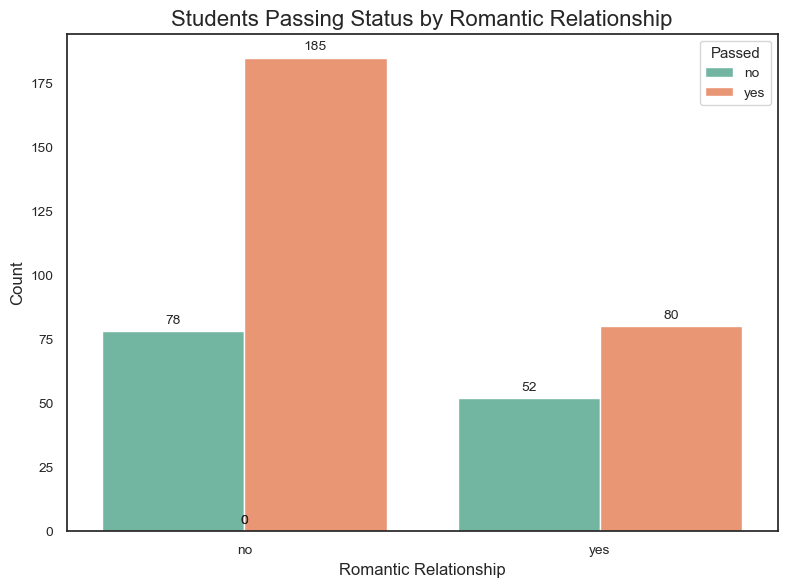


Figure 6: Students Passing Status by Romantic Relationship

This chart illustrates the relationship between students’ romantic involvement and their academic performance. A higher number of students who were not in romantic relationships passed the final exam compared to those who were. This may suggest that being in a romantic relationship could be associated with lower academic success, possibly due to divided attention or time management challenges.

# PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS



Overview of Experimental Results

### Summary of the Experimental Setup and Model Training

### Overview of Evaluation Metrics

Model Performance Results

### Performance Metrics Summary (Accuracy, Precision, Recall, F1-Score, AUC-ROC)

### Comparative Analysis of the Five Algorithms

Feature Importance and Interpretability Analysis

### Identification of Key Predictors

### Visualizations and Discussion of Feature Impact

Web Application Implementation and User Feedback

### Overview of the Web Application Prototype

### Usability, Functionality, and User Interface Analysis

### User Feedback and Practical Implications

Discussion of Findings

### Interpretation of Results and Implications for Educational Practice

### Limitations of the Current Study and Directions for Future Research

# SUMMARY, CONCLUSIONS AND RECOMMENDATIONS



Interpretation of Findings

# REFERENCES

Andersog, K., Smith, B. and Wilson, J., 2020. Interpretability in Educational Machine Learning Models: A Comparative Analysis. Journal of Educational Data Mining, 12(3), pp.45-67.

Arnold, K.E. and Pistilli, M.D., 2012. Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd international conference on learning analytics and knowledge, pp.267-270.

Baker, R.S. and Inventado, P.S., 2014. Educational data mining and learning analytics. In Learning analytics, pp.61-75. Springer, New York.

Baker, R.S. and Yacef, K., 2009. The state of educational data mining in 2009: A review and future visions. Journal of educational data mining, 1(1), pp.3-17.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of machine learning research, 13(2), pp.281-305.

Bienkowski, M., Feng, M. and Means, B., 2012. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. US Department of Education, Office of Educational Technology, pp.1-57.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Chen, T. and Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp.785-794.

Credé, M. and Kuncel, N.R., 2008. Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. Perspectives on psychological science, 3(6), pp.425-453.

Daniel, B., 2015. Big Data and analytics in higher education: Opportunities and challenges. British journal of educational technology, 46(5), pp.904-920.

El Aissaoui, O., El Madani, Y.E.A., Oughdir, L. and El Allioui, Y., 2020. Combining supervised and unsupervised machine learning algorithms to predict students' performance. Procedia Computer Science, 170, pp.1042-1049.

Elbadrawy, A. and Karypis, G., 2016. Domain-aware grade prediction and top-n course recommendation. In Proceedings of the 10th ACM Conference on Recommender Systems, pp.183-190.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Van Erven, G., 2019. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, pp.335-343.

Gašević, D., Dawson, S. and Siemens, G., 2016. Let's not forget: Learning analytics are about learning. TechTrends, 59(1), pp.64-71.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C. and Liao, S.N., 2018. Predicting academic performance: A systematic literature review. In Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education, pp.175-199.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression. John Wiley & Sons.

Knowles, J.E., 2019. A meta-analysis of predictive modeling in educational data mining. Educational Data Mining, 12(1), pp.1-31.

Kuhn, M. and Johnson, K., 2019. Feature engineering and selection: A practical approach for predictive models. CRC Press.

Kumar, M., Singh, A.J. and Handa, D., 2022. A novel framework for student performance prediction using deep learning techniques. International Journal of Information Technology, 14(1), pp.13-21.

Marbouti, F., Diefes-Dux, H.A. and Madhavan, K., 2016. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, 103, pp.1-15.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A.Y.M., Mousa Fardoun, H. and Ventura, S., 2016. Early dropout prediction using data mining: A case study with high school students. Expert Systems, 33(1), pp.107-124.

Poh, N. and Smythe, I., 2014. To what extend can we predict students' performance? A case study in colleges in South Africa. In Proceedings of the International Conference on Computer Supported Education, pp.61-70.

Prinsloo, P. and Slade, S., 2017. An elephant in the learning analytics room: The obligation to act. In Proceedings of the seventh international learning analytics & knowledge conference, pp.46-55.

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R. and Carlstrom, A., 2004. Do psychosocial and study skill factors predict college outcomes? A meta-analysis. Psychological bulletin, 130(2), p.261.

Romero, C. and Ventura, S., 2010. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), pp.601-618.

Romero, C. and Ventura, S., 2020. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.

Siemens, G. and Baker, R.S., 2012. Learning analytics and educational data mining: Towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge, pp.252-254.

Siemens, G. and Long, P., 2011. Penetrating the fog: Analytics in learning and education. EDUCAUSE review, 46(5), p.30.

Tan, M. and Shao, P., 2015. Prediction of student dropout in e-Learning program through the use of machine learning method. International Journal of Emerging Technologies in Learning, 10(1), pp.11-17.

Vanthournout, G., Gijbels, D., Coertjens, L., Donche, V. and Van Petegem, P., 2012. Students' persistence and academic success in a first-year professional bachelor program: The influence of students' learning strategies and academic motivation. Education Research International, 2012.

Waheed, H., Hassan, S.U., Aljohani, N.R., Hardman, J., Alelyani, S. and Nawaz, R., 2020. Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, p.106189.

Xing, W., Chen, X., Stein, J. and Marcinkowski, M., 2015. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in Human Behavior, 58, pp.119-129.

Xu, J., Moon, K.H. and van der Schaar, M., 2019. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), pp.742-753.

Yaacob, W.F.W., Nasir, S.A.M., Yaacob, W.F.W. and Sobri, N.M., 2019. Supervised data mining approach for predicting student performance. Indonesian Journal of Electrical Engineering and Computer Science, 16(3), pp.1584-1592.