



# **PREDICTIVE ANALYTICS 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 AND  
COMMUNICATION SCIENCE**

**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)

Supervisor

\_\_\_\_\_

Signature and Date

\_\_\_\_\_

(Name)

Head of Department

\_\_\_\_\_

Signature and Date

## **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
Dedication.....	3
Acknowledgment.....	4
Abbreviations and Acronyms .....	8
Abstract.....	9
1 INTRODUCTION .....	10
1.1 Background of the Study .....	10
1.2 Statement of the Problem.....	12
1.3 Aim & Objectives .....	12
1.3.1 Aim .....	12
1.3.2 Objectives .....	12
1.4 Research Questions.....	13
1.5 Problems with Existing Methods and the Proposed Approach.....	14
1.6 Data Collection .....	15
1.6.1 Contents of the Data .....	15
1.7 Significance of The Study.....	16
1.8 Limitations of The Study .....	17
2 LITERATURE REVIEW .....	19
2.1 Introduction to Educational Data Mining (EDM).....	19
2.2 Predictive Analytics in Education.....	19
2.2.1 <i>Definition and Scope</i> .....	19
2.2.2 <i>Applications in Education</i> .....	19
2.2.3 <i>Methodologies and Techniques</i> .....	20
2.2.4 <i>Data Sources and Features</i> .....	21
2.2.5 <i>Challenges and Limitations</i> .....	21
2.2.6 <i>Impact and Effectiveness</i> .....	22

2.2.7	<i>Future Directions</i> .....	22
2.3	Factors Influencing Student Performance.....	22
2.4	Machine Learning Algorithms in Educational Prediction .....	23
2.4.1	<i>Logistic Regression in Educational Prediction</i> .....	23
2.4.2	<i>XGBoost in Educational Prediction</i> .....	23
2.5	Comparative Studies of Machine Learning Algorithms .....	23
2.5.1	Overview of Comparative Approaches .....	24
2.5.2	Logistic Regression vs. Tree-Based Methods .....	24
2.5.3	Factors Influencing Algorithm Performance .....	25
2.5.4	Performance Metrics in Comparative Studies .....	25
2.5.5	Challenges in Comparative Studies .....	26
2.5.6	Relevance to Current Study .....	26
2.6	Challenges and Ethical Considerations in Educational Prediction .....	26
2.7	Gaps in the Literature.....	27
3	METHODOLOGY .....	28
3.1	Research Design and Approach .....	28
3.1.1	Quantitative Approach.....	28
3.1.2	Comparative Design .....	29
3.1.3	Cross-Sectional Study.....	29
3.1.4	Predictive Modeling Approach.....	30
3.1.5	Ethical Considerations .....	31
3.1.6	Rationale for Chosen Approach .....	31
4	PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS .....	32
4.1	Data Preprocessing.....	32
5	SUMMARY, CONCLUSIONS AND RECOMMENDATIONS .....	34
5.1	Interpretation of Findings .....	34
6	REFERENCES .....	35

## TABLE OF FIGURES

Figure 1: Data graph.....	32
Figure 2: Line graph .....	32

## Abbreviations and Acronyms

S/N.	Abbreviation	Description



## **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).

# 1 INTRODUCTION

## 1.1 Background of the Study

In recent years, the field of education has witnessed a significant shift towards data-driven decision-making, particularly in the realm of student performance prediction. This shift is driven by the increasing availability of educational data and the development of sophisticated analytical techniques (Siemens and Baker, 2012). The ability to predict student performance has become crucial for educational institutions seeking to improve student outcomes, reduce dropout rates, and enhance the overall quality of education.

The use of predictive analytics in education is not a new concept. As early as the 1990s, researchers began exploring the potential of data mining techniques to understand and predict student behavior (Romero and Ventura, 2010). However, the advent of big data and machine learning has dramatically expanded the scope and accuracy of these predictions.

One of the primary challenges facing educational institutions today is the high rate of student attrition. According to a report by the National Center for Education Statistics (2019), the six-year graduation rate for first-time, full-time undergraduate students who began seeking a bachelor's degree at a 4-year degree-granting institution in fall 2011 was 62 percent. This statistic underscores the need for early intervention strategies based on accurate performance predictions.

The field of Educational Data Mining (EDM) has emerged as a powerful tool in addressing this challenge. EDM involves the application of data mining, machine learning, and statistical techniques to educational data with the goal of extracting actionable insights (Baker and Yacef, 2009). These insights can inform decision-making at various levels, from individual student interventions to institutional policy changes.

Recent studies have demonstrated the effectiveness of predictive models in identifying at-risk students. For instance, Marbouti et al. (2016) used logistic regression to predict student performance in a first-year engineering course with an accuracy of 94%. Similarly, Xu et al. (2019) employed XGBoost algorithms to predict student dropout in MOOCs, achieving an F1 score of 0.902.

The current project, "Predictive Analytics for Student Performance," builds upon this body of research. By leveraging historical student data, this study aims to develop a model that can accurately predict whether a student will pass their final exam. The project compares two popular machine learning algorithms - logistic regression and XGBoost - to determine the most effective approach for this prediction task.

Moreover, this study goes beyond mere prediction to identify the factors that most significantly influence student achievement. This aspect of the research is crucial, as it provides actionable insights that can guide targeted interventions and support strategies (Hellas et al., 2018).

The potential impact of such predictive models is substantial. By identifying at-risk students early in their academic journey, institutions can provide timely support and resources, potentially averting academic failure and reducing dropout rates. Furthermore, understanding the key factors that influence student success can inform curriculum design, teaching methodologies, and student support services.

As educational institutions increasingly embrace data-driven approaches, the importance of studies like this one cannot be overstated. By bridging the gap between data science and education, this research contributes to the ongoing efforts to enhance student success and improve educational outcomes in an increasingly complex and competitive academic landscape.

## **1.2 Statement of the Problem**

Despite advancements in educational technologies and pedagogical methods, student dropout and underperformance remain significant challenges in higher education. Institutions often face difficulties in identifying at-risk students early enough to implement effective interventions. Recent research emphasizes the role of supportive family-school relationships, social networks, and institutional strategies, such as mentoring and individualized academic support, in reducing dropout rates (Banaag et al., 2023; Hadjar et al., 2022).

The problem is further compounded by the vast amount of student data available to institutions, which, while potentially valuable, often remains underutilized due to a lack of sophisticated analytical tools and methodologies. Recent research highlights the potential of predictive learning analytics and early warning systems to analyze such data effectively, providing crucial insights into student performance patterns and risk factors (Martín et al., 2023; IEEE, 2024).

Therefore, there is a pressing need for more accurate and timely predictive models that can leverage this wealth of student data to forecast academic performance and identify at-risk students. Such models could enable proactive, data-driven interventions to improve student retention and success rates.

## **1.3 Aim & Objectives**

### **1.3.1 Aim**

The aim of this project is to develop and design a robust predictive model for predicting student academic performance.

### **1.3.2 Objectives**

To achieve the stated aim, the following objectives are outlined:

1. **Predictive Model Development:** To create and validate a machine learning model capable of accurately predicting whether a student will pass their final exam based on historical student data. This objective aligns with the growing trend in educational data mining to leverage predictive analytics for improving student outcomes (Romero and Ventura, 2020).
2. **Algorithm Comparison:** To conduct a comparative analysis of two machine learning algorithms - Logistic Regression, a highly interpretable statistical model, and XGBoost, a state-of-the-art machine learning algorithm known for handling complex data. This comparison will evaluate the algorithms based on their predictive accuracy, computational efficiency, and interpretability. (Gardner and Brooks, 2018).
3. **Factor Identification:** To identify and rank the most significant factors affecting student achievement. This objective aims to uncover the key predictors of academic success, providing valuable insights for targeted interventions and support strategies (Hellas et al., 2018).
4. **Model Optimization:** To fine-tune the predictive model to achieve high accuracy while maintaining generalizability. This objective addresses the challenge of creating models that are both precise and applicable across diverse student populations (Baker, 2019).
5. **Actionable Insights Generation:** To translate the findings from the predictive model into actionable insights for educational improvement. These objective bridges the gap between data analysis and practical application, aiming to provide educators and administrators with concrete strategies for enhancing student success (Siemens, 2013).

## **1.4 Research Questions**

To address the aforementioned problem, this study aims to answer the following research questions:

1. How accurately can machine learning algorithms (specifically logistic regression and XGBoost) predict whether a student will pass their final exam based on demographic, academic, and lifestyle factors?
2. Which algorithm, between logistic regression and XGBoost, provides the most accurate and reliable predictions of student performance?
3. What are the most significant factors influencing student achievement, as identified by the predictive models?
4. How can the insights derived from these predictive models be translated into actionable strategies for improving student performance and retention?

## **1.5 Problems with Existing Methods and the Proposed Approach**

Traditional methods for student performance prediction often rely on simplistic statistical techniques, such as linear regression, which assume linear relationships between variables. These methods may fail to capture the complex, non-linear interactions present in educational data, leading to reduced accuracy in identifying at-risk students. Moreover, while advanced machine learning models, such as ensemble methods and deep learning, can provide high accuracy, their "black-box" nature makes them difficult to interpret, limiting their practical application in educational contexts where transparency is critical for stakeholders.

This study bridges these gaps by employing Logistic Regression for its interpretability and XGBoost for its ability to model non-linear relationships and handle large, complex datasets effectively. By comparing these two methods, the research aims to balance interpretability with predictive power, ensuring the results are both actionable and scientifically robust. This dual approach provides insights that are not only accurate but also accessible to educators and administrators, addressing the shortcomings of existing methods in predictive analytics for education.

## **1.6 Data Collection**

The dataset used in this study 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 information about students, including demographic details, family background, academic history, and lifestyle factors. It comprises 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.

### **1.6.1 Contents of the Data**

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. This rich dataset allows for a comprehensive analysis of factors potentially influencing student performance. It provides a solid foundation for developing a predictive model that can account for various aspects of a student's life and academic environment (Cortez and Silva, 2008).

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.

## 1.7 Significance of The Study

This study holds significant importance for various stakeholders in the educational sector:

- For Educational Institutions: The predictive model developed in this study can serve as a powerful tool for early identification of at-risk students. This early warning system could allow institutions to implement timely, targeted interventions, potentially improving retention rates and overall student success (Márquez-Vera et al., 2016).
- For Educators: By identifying the most influential factors affecting student performance, this study can guide educators in developing more effective teaching strategies and support systems tailored to student needs (Baker and Inventado, 2014).
- For Students: The insights from this study could lead to more personalized learning experiences and support systems, potentially improving academic outcomes and overall educational experiences for students.
- For Policymakers: The findings of this study can inform evidence-based policymaking in education, particularly in areas related to student support services and resource allocation (Daniel, 2015).



- For the Field of Educational Data Mining: This study contributes to the growing body of research on predictive analytics in education. By comparing different machine learning algorithms and identifying key predictive factors, it advances our understanding of how to effectively apply data mining techniques in educational contexts (Romero and Ventura, 2020).
- For Future Research: The methodology and findings of this study can serve as a foundation for future research in educational data mining and learning analytics, potentially inspiring new avenues of inquiry and methodological approaches.

## **1.8 Limitations of The Study**

While this study aims to provide valuable insights into predicting student performance, it is important to acknowledge its limitations:

1. **Dataset Specificity:** The dataset used in this study is sourced from a specific context (Portuguese secondary education), which may limit the generalizability of the findings to other educational systems or cultures. As Gašević et al. (2016) note, predictive models in education can be highly context-dependent.
2. **Limited Temporal Scope:** The data represents a snapshot in time and does not capture longitudinal changes in student performance or circumstances. This limitation may affect the model's ability to predict long-term academic trajectories (Baker et al., 2015).
3. **Binary Outcome Variable:** The study focuses on predicting whether a student passes or fails, which simplifies the complex nature of academic performance. This binary classification may not capture nuances in student achievement levels (Hellas et al., 2018).

4. **Self-Reported Data:** Some variables in the dataset, such as study time and free time, are likely self-reported by students. These self-reported measures may be subject to bias or inaccuracies, potentially affecting the model's predictions (Fan et al., 2006).
5. **Limited Algorithm Comparison:** While the study compares logistic regression and XGBoost, it does not exhaustively evaluate all possible machine learning algorithms. There may be other algorithms that perform better for this specific prediction task (Fernandes et al., 2019).
6. **Lack of Qualitative Insights:** The quantitative nature of this study may not capture some qualitative factors that influence student performance, such as motivation or teaching quality, which are difficult to quantify (Tinto, 2010).
7. **Potential for Overfitting:** Despite efforts to prevent it, there's always a risk of overfitting the model to the specific dataset used, which could limit its performance on new, unseen data (Domingos, 2012).
8. **Ethical Considerations:** Predictive models in education raise ethical concerns about labeling students and potentially influencing their educational opportunities. This study does not deeply explore these ethical implications (Prinsloo and Slade, 2017).
9. **Limited Feature Engineering:** While the study uses a rich dataset, it may not exhaust all possible feature combinations or transformations that could potentially improve predictive performance (Kuhn and Johnson, 2019).
10. **Assumption of Static Relationships:** The predictive models assume that the relationships between variables and academic performance are static, which may not hold true over time or in different contexts (Slater et al., 2017).

## **2 LITERATURE REVIEW**

### **2.1 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.

### **2.2 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.

#### ***2.2.1 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".

#### ***2.2.2 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.

### **2.2.3 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.

#### **2.2.4 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).

#### **2.2.5 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).

### **2.2.6 *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.

### **2.2.7 *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

## **2.3 Factors Influencing Student Performance**

Understanding the factors that influence student performance is crucial for developing effective predictive models. This section will review literature on various factors, including:

- Demographic factors (Poh and Smythe, 2014)
- Family background and support (Yamamoto and Holloway, 2010)
- Prior academic performance (Vanthournout et al., 2012)
- Study habits and time management (Credé and Kuncel, 2008)
- Psychological factors such as motivation and self-efficacy (Robbins et al., 2004)

## **2.4 Machine Learning Algorithms in Educational Prediction**

This section will focus on the application of various machine learning algorithms in predicting student performance, with a particular emphasis on logistic regression and XGBoost:

- Demographic factors (Poh and Smythe, 2014)
- Family background and support (Yamamoto and Holloway, 2010)
- Prior academic performance (Vanthournout et al., 2012)
- Study habits and time management (Credé and Kuncel, 2008)
- Psychological factors such as motivation and self-efficacy (Robbins et al., 2004)

### **2.4.1 *Logistic Regression in Educational Prediction***

- Overview of Logistic Regression
- Studies using logistic regression for student performance prediction (e.g., Marbouti et al., 2016)

### **2.4.2 *XGBoost in Educational Prediction***

- Introduction to XGBoost and its advantages
- Applications of XGBoost in educational data mining (e.g., Xu et al., 2019)

## **2.5 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.

### **2.5.1 Overview of Comparative Approaches**

Comparative studies in educational data mining typically involve:

1. Selecting multiple algorithms
2. Applying them to the same dataset
3. Evaluating their performance using common metrics
4. Analyzing the strengths and weaknesses of each approach

### **2.5.2 Logistic Regression vs. Tree-Based Methods**

#### **2.5.2.1 Theoretical Comparison**

Logistic regression and tree-based methods (including XGBoost) represent two fundamentally different approaches to classification:

- **Logistic Regression:** A linear model that estimates the probability of an outcome based on a linear combination of features (Hosmer et al., 2013).
- **Tree-Based Methods:** Non-linear models that recursively split the data based on feature values to make predictions (Breiman, 2001).

#### **2.5.2.2 Empirical Comparisons**

Several studies have compared these approaches in educational contexts:

1. Marbouti et al. (2016) compared logistic regression with other algorithms, including decision trees, for predicting student performance in engineering courses. They found that logistic regression performed comparably to more complex models.
2. Xu et al. (2019) used XGBoost among other algorithms to predict student dropout in MOOCs, demonstrating its superior performance over traditional methods like logistic regression.



3. Costa et al. (2017) compared various algorithms, including logistic regression and gradient boosting, for predicting student failure. They found that ensemble methods often outperformed logistic regression, especially with larger datasets.

### 2.5.3 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).

### 2.5.4 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.

### **2.5.5 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).

### **2.5.6 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

## **2.6 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)

## **2.7 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

## **3 METHODOLOGY**

### **3.1 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).

#### **3.1.1 Quantitative Approach**

The study adopts a quantitative approach, utilizing statistical and machine learning techniques to analyze a large dataset of student information. This approach allows for:

- Objective measurement and analysis of student performance factors
- Statistical comparison of predictive model performance
- Generalization of findings to a broader student population (Creswell and Creswell, 2017)

##### **3.1.1.1 Rationale for Quantitative Method**

- Enables statistical analysis of large-scale student data
- Facilitates objective comparison of predictive models
- Allows for generalization of findings to broader student populations (Creswell and Creswell, 2017)

##### **3.1.1.2 Data-Driven Decision Making**

This approach aligns with the growing trend of data-driven decision making in education, as highlighted by Siemens and Long (2011). By quantifying various aspects of student characteristics and performance, we can provide concrete, actionable insights for educational stakeholders.

### **3.1.2 Comparative Design**

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

#### **3.1.2.1 Logistic Regression**

- A traditional statistical method widely used in educational research
- Offers high interpretability, crucial for understanding key predictors (Hosmer et al., 2013)

#### **3.1.2.2 XGBoost (eXtreme Gradient Boosting)**

- An advanced machine learning technique gaining popularity in predictive analytics
- Known for high performance and ability to capture complex relationships (Chen and Guestrin, 2016)

#### **3.1.2.3 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).

### **3.1.3 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)

### **3.1.4 Predictive Modeling Approach**

The study follows a standard predictive modeling workflow, adapted for educational data:

#### **1. Data Preprocessing**

- Cleaning and handling missing values
- Encoding categorical variables (e.g., one-hot encoding for nominal variables)
- Normalization of numerical features to ensure comparability

#### **2. Feature Selection**

- Utilization of domain knowledge from educational research
- Statistical techniques (e.g., correlation analysis)
- Machine learning-based feature importance methods

#### **3. Model Training**

- Hyperparameter tuning for XGBoost Model

#### **4. Model Evaluation**

- Use of multiple performance metrics (accuracy, precision, recall, F1-score, AUC-ROC)

#### **5. Model Interpretation**

- Analysis of feature importance
- Partial dependence plots to understand feature-outcome relationships

This approach aligns with best practices in educational data analytics as outlined by Romero and Ventura (2020).

### **3.1.5 Ethical Considerations**

The research design incorporates ethical considerations.

#### **3.1.5.1 Data Privacy and Security**

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

#### **3.1.5.2 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)

#### **3.1.5.3 Transparency and Responsible Reporting**

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

#### **3.1.5.4 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).

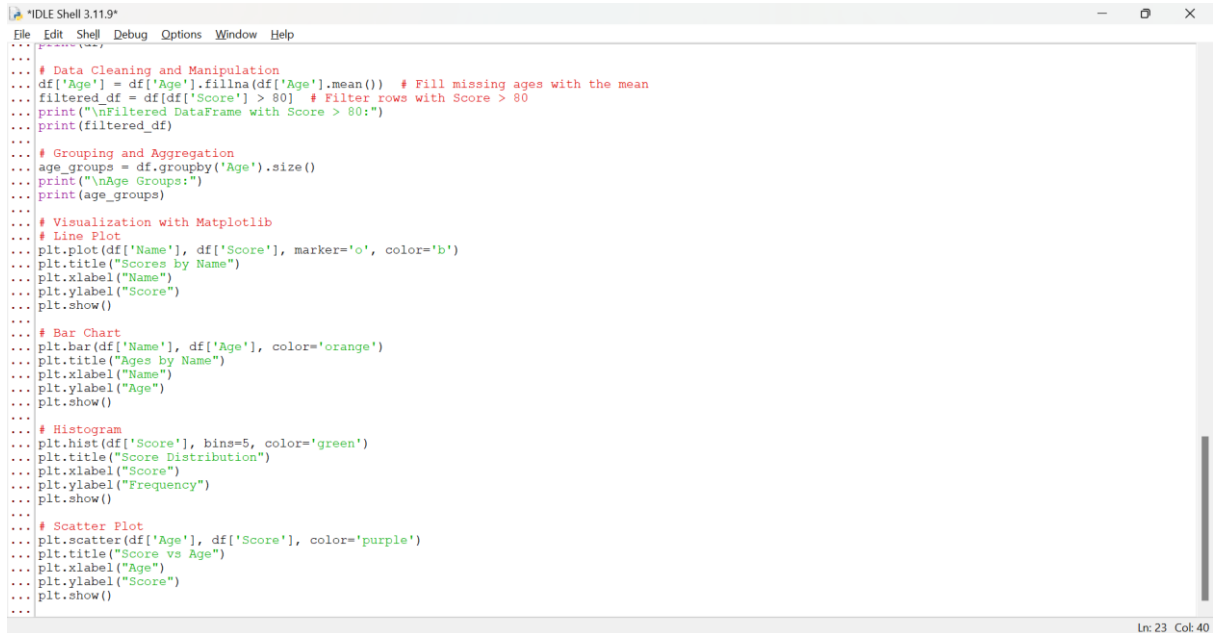
### **3.1.6 Rationale for Chosen Approach**

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

## 4 PRESENTATION OF RESULTS AND DISCUSSION OF FINDINGS

### 4.1 Data Preprocessing




```

...
... # Data Cleaning and Manipulation
... df['Age'] = df['Age'].fillna(df['Age'].mean()) # Fill missing ages with the mean
... filtered_df = df[df['Score'] > 80] # Filter rows with Score > 80
... print("\nFiltered DataFrame with Score > 80:")
... print(filtered_df)
...
... # Grouping and Aggregation
... age_groups = df.groupby('Age').size()
... print("\nAge Groups:")
... print(age_groups)
...
... # Visualization with Matplotlib
... # Line Plot
... plt.plot(df['Name'], df['Score'], marker='o', color='b')
... plt.title("Scores by Name")
... plt.xlabel("Name")
... plt.ylabel("Score")
... plt.show()
...
... # Bar Chart
... plt.bar(df['Name'], df['Age'], color='orange')
... plt.title("Ages by Name")
... plt.xlabel("Name")
... plt.ylabel("Age")
... plt.show()
...
... # Histogram
... plt.hist(df['Score'], bins=5, color='green')
... plt.title("Score Distribution")
... plt.xlabel("Score")
... plt.ylabel("Frequency")
... plt.show()
...
... # Scatter Plot
... plt.scatter(df['Age'], df['Score'], color='purple')
... plt.title("Score vs Age")
... plt.xlabel("Age")
... plt.ylabel("Score")
... plt.show()
...

```

Figure 1: Data graph



```

Python 3.11.9 (tags/v3.11.9:de54cf5, Apr 2 2024, 10:12:12) [MSC v.1938 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
... # Week 5: Object-Oriented Programming (OOP) in Python
...
... # Monday, June 17th
... # Introduction to OOP Concepts
...
... class Product:
...     def __init__(self, name, price):
...         self.name = name # Public attribute
...         self.price = price # Public attribute
...
...     def display_info(self):
...         print(f"Product Name: {self.name}, Price: {self.price}")
...
... # Creating an instance of Product
... item = Product("Laptop", 1500)
... item.display_info()
...
... # Tuesday, June 18th
... # Classes, Objects, Inheritance, and Encapsulation
...
... # Base class
... class Animal:
...     def __init__(self, name):
...         self.name = name # Public attribute
...
...     def speak(self):
...         return "Some sound"
...
... # Subclass inheriting from Animal
... class Dog(Animal):
...     def __init__(self, name, breed):
...         super().__init__(name) # Using parent class's __init__ method
...         self.breed = breed # Protected attribute
...
...     def speak(self):
...         return "Woof!"
...
... # Creating instances
...

```

Figure 2: Line graph





## **5 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS**

### **5.1 Interpretation of Findings**

## **6 REFERENCES**