

Predicting Student Performance Using Ensemble Learning: An Xgboost Approach with Model Evaluation and Insight

Abdulgaffar Abubakar Yahaya^{1*}; Muktar Danlami, Ph.D.²; Zawali S. Paki²; Hamza Aliyu Idris³; Usman Haruna⁴; Ismail Abubakar Yahaya⁵

¹Cyber Security/Computing/Northwest University.

²Northwest University, Kano.

³Computing, Bayero University, Kano.

⁴Computing, Mewar University.

⁵Computing

*Corresponding Author's Email: abdulgaffarabubakar40@gmail.com

Abstract

Predicting student academic performance was important for identifying students at risk of poor achievement and supporting timely educational interventions. This study developed and evaluated an XGBoost-based machine learning model for predicting student performance using academic, behavioural, and socioeconomic factors. The study used a dataset obtained from the Kaggle repository containing 6,607 student records and 20 features. The data were preprocessed to address missing values and prepare categorical and numerical variables for modelling. An XGBoost regression model was trained using 80% of the dataset and evaluated on the remaining 20% using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). The model achieved an MSE of 4.75, an MAE of 0.97, and an R^2 of 0.66, indicating that it explained approximately 66% of the variation in student performance. Feature-importance analysis showed that hours studied, previous scores, and attendance were the most influential predictors, while sleep, tutoring, parental involvement, and socioeconomic factors also contributed to prediction. However, the model showed greater difficulty predicting students at the higher end of the performance distribution. The findings demonstrated the usefulness of XGBoost for student performance prediction while highlighting the need for richer features, balanced datasets, and further model optimization.

Keywords: Student Performance Prediction; XGBoost; Ensemble Learning; Machine Learning; Educational Data Mining; Predictive Analytics; Student Academic Performance.

How to Cite: Yahaya, A. A., Danlami, M., Paki, Z. S., Idris, H. A., Haruna, U., & Yahaya, I. A. (2026). Predicting Student Performance Using Ensemble Learning: An Xgboost Approach with Model Evaluation and Insight. British Journal of Computer, Networking and Information Technology, 9(3), 1–18. <https://doi.org/10.52589/BJCNIT-VRC734FZ>

INTRODUCTION

Among the many different observable phenomena in the students' careers, University dropout is one of the most complex and adverse events, both for students and institutions (Del Bonifro et al., 2020). Student dropout is a dangerous situation that affects the student both emotionally and economically. It also causes a great setback, most especially to the institutions and the community in general. Therefore, academic success and failure are very vital factors for economic success in the knowledge-based economy (Yakubu & Abubakar, 2022).

Artificial Intelligence (AI) is changing many aspects of our society and our lives since it provides the technological basis for new services and tools that help decision-making in everyday life (Del Bonifro et al., 2020). Nowadays, technology has become an indispensable tool that transforms our lives and experiences. Artificial intelligence and machine learning have great significance in the field of education, as they brought new possible ways of improving the teaching and learning process. An important aspect is the ability to transform the education field by the development of a model that predicts the students' academic success and failure. This tool will help the students to improve their academic careers.

This research embarks on a journey of using machine learning algorithms, specifically XGBoost, to develop a predictive model for education. A large amounts of data is collected each year by the educational institutions, including information regarding the academic path of the student, as well as demographics and socio-economic factors. This model is going to use that data to predict the student performance in the educational institutions. The research leverages the use of the XGBoost algorithms to perform the prediction, and it aims to answer the following questions:

1. What factors contribute to the student performance in academic institutions?
2. What are the key predictors of student success within these institutions, and how do they influence outcomes?
3. How are the predictive capabilities of XGBoost regression in forecasting student performance?

This research work is organized as follows: part 1 presents an introduction to the work, while part 2 contains a comprehensive review of existing research on dropout rates, student success, and the application of machine learning in education. Part 3 describes the methodology. A detailed explanation of the data collection process, including a description of the data, the methods used to deal with the unbalanced dataset, and the procedures for training and evaluation of the classification models. Part 4 contains presentation and interpretation of findings derived from the machine learning models. Part 5 discusses the implications of the results, limitations of the study, and avenues for future research.

LITERATURE REVIEW

Academic success prediction is not a novel concept in education, as several scholars have attempted different kinds of methods utilizing various models and data. Researchers and the stakeholders in higher education institutions are very interested in determining whether a student is likely to drop out or not finish their degree and this knowledge might be useful in developing methods to support such individuals (Lone & Wani, 2022)

Several papers that utilized the machine learning algorithms for the prediction of students' academic performance and dropout rate had been surveyed to come up with the detailed review of the literature.

(Martins et al., 2021) Develop classification models to predict the academic performance of students at the time of enrollment in a higher education institution. Utilize enrollment-time data, including demographics, prior academic performance, and socio-economic factors. Construct models using both standard machine learning algorithms and boosting algorithms. Boosting algorithms outperform standard methods for this classification task but still struggle to accurately classify minority classes.

(Yakubu & Abubakar, 2022) study aims to predict students' academic performance using early detection indicators to facilitate timely remediation, thereby enhancing economic success in a knowledge-based economy. They make use of a machine learning approach to develop a predictive model. Utilize secondary data from a Nigerian university's student information system. Analyze factors such as age, gender, high school exam scores, region, and CGPA. Findings show that Female students are 1.2 times more likely to achieve a high CGPA compared to male students. Students from affluent and developed regions or those in Years 3 and 4 are more likely to achieve a high CGPA. (Del Bonifro et al., 2020) analyse a dataset of 15000 students enrolled in different courses of academic year 2016/2017 and developed a machine learning classifier model using the support vector machine (SVM), random forest (RF), and Linear discriminant analysis. The model can be used to predict the dropout of the first-year undergraduate students and to identify the risk of quitting an academic course in the first year or during the student application process.

(Vijayalakshmi & Venkatachalapathy, 2019) using a dataset of 500 students obtained from the Kaggle Machine learning repository. The dataset has 16 features that are grouped into 3 classes, such as Demographic, academic background, section, and behavioural features. They employed the use of six machine learning techniques, such as Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, K-Nearest Neighbor, and Deep neural network in R Programming and compared the accuracy of all other algorithms. Among six algorithms, it is shown that Deep Neural Network outperformed with 84% accuracy.

(Ajibade et al., 2019) Analyses a dataset containing a collection of 500 students with 16 features that is obtained from a learning management system called Kalboard 360. They make use of machine learning algorithms such as support vector machine (SVM), K-Nearest Neighbour (KNN), decision Tree (DT) and ensemble methods to propose a new performance prediction model for students based on data mining methods.

(Mduma et al., 2019) Provide an overview about the different machine learning algorithms that can be used in an educational setting. They focused their discussion on identifying the techniques that can be used to predict student dropout in higher institutions.

(Singh & Pal, 2020) Developed a machine learning model to enhance the prediction of student performance using various ML algorithms such as Decision Tree (DT), Naïve Bayesian (NB), K-Nearest Neighbours (KNN), Extra Tree (ET), and ensemble techniques. The model was investigated using a student dataset consisting of 1000 instances and 22 attributes. The result obtained with each ML algorithm was compared and tested with various factors like accuracy, sensitivity, specificity, and F1-score. Result after the comparison shows that bagging is the best method, which gives a better result as compared to bagging ensemble techniques. The developed model can be applied to the admission-seeking students to identify the prediction of

their performance in the selected course, which can be beneficial for both the students and Institution.

(Evangelista & Sy, 2022) Explores how to improve the accuracy of predicting student performance in online courses. Using the standard machine learning algorithms such as multilayer perceptron or neural networks (NN), random forest (RF), naïve Bayes (NB), J48, JRip, OneR, logistic regression (LR), k-nearest neighbour (KNN), and support vector machine (SVM), they developed and tested a model to predict performance. Their research focuses on a technique called ensemble learning that combines multiple learning algorithms to create a more powerful prediction model. They make use of the homogeneous ensembles (bagging and boosting) and heterogeneous ensembles (voting and stacking). Using the publicly available dataset of student performance from the University of California Irvine (UCI) machine learning repository. The comparative analysis of the model's accuracy showed that the best-performing single classifier's accuracy increased further from 93.10% to 93.68% when used as a base classifier of a voting ensemble method. Moreover, results in this study showed that voting heterogeneous ensemble performed slightly better than bagging and boosting homogeneous ensemble methods.

(Ahammad et al., 2021) Implement a model to predict student performance based on their academic results in the SSC exam. By using a dataset that contains a collection of 400 students, the model compares different machine learning algorithms to build machine learning models that will be used for the prediction task. Five machine learning algorithms were evaluated: Naive Bayes, K-Nearest Neighbors, Support Vector Machine, XGBoost, and Multi-Layer Perceptron (MLP). Data preprocessing steps are performed to ensure data quality before feeding it into the model. The final result shows that all five algorithms achieved satisfactory results, with accuracy exceeding 80%, but the MLP model achieved the highest accuracy of 86.25% in predicting student exam results.

(Berens et al., 2018) Addresses the issue of students' dropout in German universities by developing an Early Detection System (EDS) that predicts any student at risk in the university. The system can identify students at risk of dropping out early so universities can intervene and potentially improve retention rates. In training the model, they made use of administrative data from German universities that included demographics and academic performance and employed a machine learning technique called AdaBoost to combine multiple prediction methods. The EDS successfully predicts student dropout with increasing accuracy over time.

(Sonparote et al., 2023) Compare different Educational data Mining (EDM)/Learning Analytics (LA) techniques to identify their potential strengths and weaknesses that are applied in the educational domain to predict the student's performance. They try to answer 2 questions in this work as follows: (i) What EDM/LA techniques are used for predicting student performance? And (ii) which machine learning technique performs better by comparing them in terms of strength, weakness, and accuracy. The result achieve shows that, it is indecisive which technique predicts the student's performance more accurately as different authors present different results. Also, most authors used supervised learning techniques, whereas only a few authors used unsupervised learning techniques for the prediction of students' performance. It is an opportunity for the researchers to conduct further research in unsupervised learning techniques. Also, 52% of the papers reviewed have predicted students' performance at the university level.

Table 1: Table of related literature

Title	Year	Author	Method	Result
Early Prediction of students' Performance in Higher Education: A Case Study	2021	Martins et al.	Ensemble approach	Boosting algorithm outperformed standard method
Applying machine learning approach to predict students' performance in higher educational institutions	2022	Yakubu and Abubakar	Logistic regression	Female students are 1.2 times more likely to achieve higher CGPA
Student dropout prediction	2020	Del Bonifro et al.	SVM, RF, and Linear Discriminant Analysis (LDA)	SVM and LDA obtain higher performance
Comparison of Predicting Students' Performance using Machine Learning Algorithms	2019	Vijayalakshmi and Venkatachalapathy	DNN, NB, RF, SVM, KNN, DT with R-Programming	DNN outperforms with 84% accuracy
Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods	2019	Ajibade et al.	Ensemble method	Ensemble methods performed with 91.5 accuracy when applied to the classifier
A survey of machine learning approach techniques for student dropout prediction	2019	Mduma et al.	Propose and analyze different ML techniques for student prediction	Result shows that most of the machine learning techniques have been tested in the developed countries
Machine Learning Algorithms and Ensemble Technique to	2020	Singh and Pal	DT, NB, KNN, ET plus ensemble learning (Bagging and	Bagging gives better results.

Improve Prediction of Students Performance		Boosting)		
An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods	2022	Evangelistic and Benedict	NN, LR, RF, NN, and ensemble method	Voting heterogeneous outperformed homogeneous bagging
A Comparative Study of Different Machine Learning Techniques to Predict the Result of an Individual Student using Previous Performances	2021	Ahammad et al.	NB, KNN, SVM, XGBoost, and MLP	MLP achieved highest accuracy of 86.25% in predicting student exams.
Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods	2018	Beren et al.	Ensemble learning AdaBoost	Successfully predict student dropout with increasing accuracy
Comparative Analysis of Various Techniques used	2021	Yashashree K. Sonparote et al.	Analyses various techniques used for predicting student performance.	It is indecisive which technique predicts the student's performance more accurately as different authors present different results. Also most authors used supervised learning techniques

Methodologically, previous research has compared conventional machine learning algorithms with ensemble approaches. Studies using Decision Trees, Naïve Bayes, K-Nearest Neighbours, Support Vector Machines, Random Forest, Neural Networks, and boosting techniques have reported varying levels of predictive performance (Vijayalakshmi & Venkatachalapathy, 2019), (Singh & Pal, 2020), (Ahammad et al., 2021). Ensemble methods have generally demonstrated strong predictive capability. (Martins et al., 2021), for instance, reported that boosting methods outperformed standard classification approaches, while (Ajibade et al., 2019) and (Evangelista & Sy, 2022) found strong performance from ensemble approaches. However, these findings do not establish a universally superior algorithm. A comparative review by (Sonparote et al., 2023) reported that different studies produced different results regarding the most accurate technique, indicating that model performance was dependent on the dataset, variables, and educational context. Thus, although ensemble learning has shown considerable promise, further empirical evaluation of specific ensemble algorithms remains necessary.

The literature also reveals important differences in research context and data characteristics. Several studies relied on relatively small datasets, such as datasets containing 400–1,000 students, while others focused on particular institutional settings or specific forms of student data (Vijayalakshmi & Venkatachalapathy, 2019), (Singh & Pal, 2020), (Ahammad et al., 2021). Research on dropout prediction has also demonstrated the usefulness of administrative and institutional data for identifying students at risk (Del Bonifro et al., 2020), (Berens et al., 2018). Nevertheless, the findings across studies are difficult to generalize because differences in sample size, predictor variables, educational environments, and outcome definitions can substantially affect model performance. Moreover, the existing literature has predominantly emphasized supervised learning, with comparatively limited attention given to alternative approaches (Sonparote et al., 2023).

Another important issue concerns the treatment of student performance as a multidimensional outcome. Although previous studies have successfully predicted academic performance or dropout, many have concentrated on classification accuracy rather than examining how different academic, behavioural, health, and socioeconomic characteristics jointly contribute to predicted performance. The literature therefore provides evidence that factors such as previous performance, attendance, study behaviour, and socioeconomic background can be useful predictors, but there remains a need to determine which factors are most influential within a combined predictive framework. This is particularly relevant because studies have produced inconsistent conclusions about the relative effectiveness of different algorithms and predictor sets (Singh & Pal, 2020), (Evangelista & Sy, 2022), (Sonparote et al., 2023).

A further gap concerns predictive performance at the extremes of the student-performance distribution. Existing research has generally emphasized overall predictive accuracy, but comparatively less attention has been given to whether models perform equally well for low-, average-, and high-performing students. This issue is important because imbalanced outcome distributions can cause predictive models to perform well for the majority group while producing less reliable predictions for minority groups. The present study encountered this limitation, as the dataset contained an uneven distribution of high-, average-, and low-performing students and the model showed difficulty predicting students at the performance extremes.

Therefore, the principal research gap lies in the limited and inconsistent evidence regarding the effectiveness of XGBoost for predicting student performance using a combination of

academic, behavioural, health, and socioeconomic characteristics, particularly when attention is given not only to overall predictive performance but also to the relative importance of individual predictors and prediction errors among high-performing students. Although previous research has established the usefulness of machine learning and ensemble methods, the literature has not provided a consistent conclusion about the best-performing approach or sufficiently examined the factors that explain prediction difficulties across different performance levels. This study addressed this gap by applying XGBoost to a dataset containing 6,607 student records and 20 features, evaluating its predictive performance using MSE, MAE, and R^2 , and examining feature importance to identify the factors that contributed most strongly to student performance.

METHODOLOGY

This part will discuss the methodological design of the paper. As stated before, the aim of this study is to analyse student performance in educational institutions and, by using Xtreme Gradient boost algorithm (XGBoost), an ensemble method, to identify key predictors and improve predictive accuracy.

Libraries and Dependencies

To begin the analysis, necessary libraries were imported to facilitate data handling, preprocessing, model training, evaluation, and visualization. The Pandas library was used for data manipulation and analysis in a tabular format, while Numpy was employed for numerical computations. Visualization was handled through Matplotlib and Seaborn, which enabled graphical representation of data distributions and model performance. The XGBoost algorithm, known for its efficiency and accuracy, was imported from the xgboost library. Additionally, metrics from Sklearn such as accuracy score, confusion matrix, and ROC curve were essential for evaluating the model.

This initial step ensures that all required tools are available for the subsequent steps in the analysis.

Loading and Exploring the Dataset

The dataset, containing information on student performance, was loaded into the environment using Pandas' `read_csv()` function. An initial inspection of the dataset was carried out using functions such as `head()`, `info()`, and `describe()` to understand the structure, data types, and basic statistical properties. This exploration helped identify features, check for missing values, and detect any inconsistencies or anomalies that might require attention. Understanding the dataset was crucial for making informed decisions during preprocessing and model building.

Handling Missing Values

Missing values can significantly impact the performance of machine learning models. Therefore, a check for missing values was conducted using `isnull().sum()`, and appropriate steps were taken to handle them. Depending on the nature and amount of missing data, strategies like mean/median imputation or removal of rows/columns were used. This step ensured that the dataset was clean and free from inconsistencies before model training. It helps to improve completeness and consistency, preventing potential model inaccuracies due to missing data.

Data Preprocessing

Data preprocessing involved transforming the dataset into a format suitable for machine learning. This step included encoding categorical variables using methods like one-hot encoding for features such as gender or education level and normalizing or scaling numerical data to ensure uniformity.

Feature Engineering

This involves selecting the features that contribute most to the training and prediction in our dataset. Feature selection was performed to retain only the most relevant features based on domain knowledge or statistical tests. Feature engineering and preprocessing play a pivotal role in enhancing model performance by ensuring that the input features are clean, consistent, and relevant to the prediction task.

Splitting the Data into Training and Test Sets

To evaluate the model's performance fairly, the dataset was split into two parts: a training set and a test set. The training set was used to fit the XGBoost model, while the test set was used to assess its generalization to unseen data. Typically, an (80%_20%) split was employed using `train_test_split()` from Sklearn, ensuring that the model was not overfitting to the training data. Therefore, the dataset was split into 80% training data and 20% test data to enable a realistic assessment of the model's performance on unseen data and help in avoiding overfitting.

Model Training with XGBoost

The XGBoost classifier was trained on the prepared training data. XGBoost is an efficient and scalable gradient boosting algorithm known for its high performance in classification tasks. The model was fit using `XGBClassifier()`, and hyperparameters like learning rate, max depth, and number of estimators were tuned to optimize its performance. This training phase involved building decision trees iteratively and adjusting the model to minimize errors. XGBoost was chosen due to its ability to handle imbalanced data, handle large datasets efficiently, and deliver high accuracy, making it ideal for predicting student performance.

Model Evaluation

The evaluation metrics we employed to assess the model's performance included Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). After training our XGBoost regression model on 80% of the dataset, we evaluated it using the remaining 20%.

The key metrics for evaluating the model's performance are:

- Mean Squared Error (MSE): This metric quantifies the average squared difference between the predicted and actual exam scores. A lower value indicates a better fit.
- Mean Absolute Error (MAE): This measures the average magnitude of errors in a set of predictions, without considering their direction. Like MSE, lower values are preferable.
- R-squared (R^2): This statistic provides an indication of how well the independent variables explain the variability of the dependent variable. Values closer to 1 suggest a strong explanatory power.

DISCUSSION

Model Performance Overview

The model was trained using an XGBoost regression model on the dataset and achieved an overall accuracy of around 66%. This accuracy measures the proportion of correct predictions made by the model out of all predictions, providing a general indication of its effectiveness. Given the nature of the dataset, which includes both behavioral and socio-economic factors, achieving this accuracy is a promising result. However, it's essential to note that accuracy alone may not fully capture the nuances of model performance, especially in cases where class distributions are imbalanced. To evaluate the performance of the XGBoost model in predicting student performance, three key metrics were calculated: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2). These metrics provide valuable insights into how well the model is doing and where there is room for improvement.

Table 2: Metrics Table

S/N	Metrics	Values
1.	Mean Squared Error (MSE)	4.75
2.	Mean Absolute Error (MAE)	0.97
3.	R-squared (R^2)	0.66

Mean Squared Error (MSE)

The Mean Squared Error of 4.75 gives us a sense of the average squared difference between the actual and predicted values. In simpler terms, it's telling us how far off the model's predictions are from the actual scores, with larger errors penalized more heavily. An MSE of 4.75 is relatively low, meaning that the model's predictions are fairly close to the actual student performance on average. However, since MSE squares the error values, it tends to give more weight to larger errors, so while the model is doing well overall, it could be significantly off in a few cases, likely for higher-achieving students, as seen in the scatter plot.

Mean Absolute Error (MAE)

The Mean Absolute Error is another way to measure how far the model's predictions are from the actual scores, but instead of squaring the differences, it simply takes the absolute value. With an MAE of 0.97, we know that on average, the model is off by less than 1 point per prediction. This is quite impressive and highlights that for most students, the model is very close to predicting their actual performance accurately. MAE is easier to interpret compared to MSE since it reflects the average size of the errors in the same units as the original data, so here, we can say that the model is, on average, about 1 point away from students' real scores.

R-Squared (R^2)

The R-squared value of 0.66 means that 66% of the variance in student performance is explained by the model. In other words, the features used to train the model (such as previous grades, attendance, or any other academic data) are able to account for around two-thirds of the variation in how students perform. While this is a solid result, it does indicate that about onethird of the variation in student performance is still unexplained by the model. There could

be additional factors like personal motivation, study habits, or family background that weren't included in the model but play a role in influencing student outcomes. A higher R-squared value would suggest a more complete understanding of student performance, so this metric gives us a clear direction for improvement, such as incorporating more comprehensive data.

Together, these metrics paint a detailed picture of the model's performance. The low MSE and MAE suggest that the model is making fairly accurate predictions overall, with most predictions being very close to the actual values. However, the R-squared score of 0.66 points out that there's still a gap in what the model can explain. While the model is effective, particularly for average-performing students, it leaves some room for improvement when it comes to predicting students at the extremes of performance. This could be seen in the scatter plot, where outliers, especially for high-performing students, show the model's challenges in making perfect predictions for everyone.

Given these results, a few improvements can be suggested. Feature engineering, finding new, meaningful features to include in the model, might help boost the R-squared value, thereby explaining more of the variability in student outcomes. For instance, adding data on extracurricular activities, social factors, or specific learning behaviors could help the model get a better sense of what drives top-performing students, potentially reducing the errors seen in outliers. Additionally, hyperparameter tuning could further refine the model, helping to improve both the R-squared value and reduce errors across the board.

These metrics show that the XGBoost model performs well, with predictions being quite close to actual values, particularly for most students. While there's room to improve the model's understanding of student performance, especially at the higher end, these results form a solid foundation for building an even more accurate predictive model in the future.

Feature Importance

The feature importance plot revealed which factors contributed the most to predicting student performance. This chart provides a deeper understanding of which factors influence student outcomes the most and how the model weighs each of them. In our case, the most important features that contribute to student performance included factors like Hours Studied, Attendance, and Previous Score. The feature importance plot shown below represents the contribution of each feature in predicting student performance using the XGBoost model.

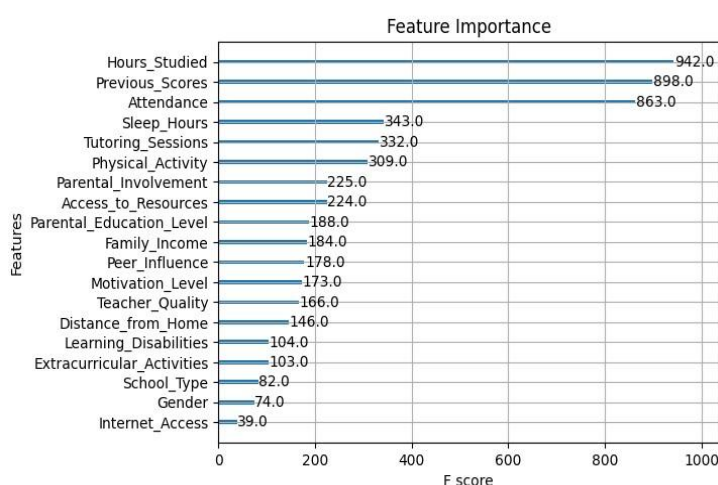


Figure 1: F-score

Key Features Of Student Performance

1. Hours Studied (F score: 942): Unsurprisingly, the number of hours a student spends studying comes out as the most influential factor. This makes sense because time spent engaging with the material is directly linked to better comprehension and retention. The fact that this feature is used most frequently in the model's decision-making highlights the importance of disciplined study habits. Students who consistently dedicate time to study are much more likely to perform better academically.
2. Previous Scores (F score: 898): Following closely behind, we see that a student's past performance plays a huge role in predicting future outcomes. This aligns with the idea that academic achievement tends to be cumulative; strong foundational knowledge leads to continued success. It's important to note that while previous performance is a strong predictor, it should not be seen as deterministic. With proper interventions (like tutoring), lower-performing students can still improve.
3. Attendance (F score: 863): This factor emphasizes the importance of being present. Regular attendance not only shows a student's commitment but also maximizes their learning opportunities. Students who attend classes regularly benefit from real-time interactions, explanations, and clarifications, which often translates into better academic performance. This finding suggests that schools and educators might focus on improving attendance to boost overall outcomes.

Health and Support-Related Factors

1. Sleep Hours (F score: 343): Interestingly, the number of hours a student sleeps is the next most important factor. This highlights the connection between good sleep hygiene and cognitive performance. Well-rested students are better able to focus, retain information, and perform in exams. This insight suggests that encouraging healthy sleep habits could have a meaningful impact on student success.
2. Tutoring Sessions (F score: 332): Tutoring plays a crucial role in supplementing regular class learning. Students who receive additional academic support tend to perform better, as tutoring offers personalized attention and addresses specific areas where a student may struggle. This insight can be leveraged by schools to provide or expand tutoring services, especially for students who are falling behind.
3. Physical Activity (F score: 309): Physical activity might seem unrelated to academic performance at first glance, but its importance here demonstrates the value of a balanced lifestyle. Exercise can improve mental clarity, reduce stress, and improve mood all of which contribute to better learning outcomes.

Social and Family Influence

1. Parental Involvement (F score: 225) and Access to Resources (F score: 224): These two factors highlight the significant role that a student's home environment plays in their academic journey. Parental involvement in education such as helping with homework, attending parent-teacher meetings, and providing guidance can have a positive influence on students' attitudes toward school and their performance. Likewise, having access to educational resources (like books, computers, or a quiet place to study) can make a substantial difference in learning outcomes. Schools and

communities could focus on enhancing these aspects to support students, particularly in underprivileged areas.

2. Parental Education Level (F score: 188) and Family Income (F score: 184): While these factors are less influential than personal effort or direct support, they still play a role. Students whose parents have higher levels of education may benefit from a home environment that values and supports academic achievement. Similarly, family income can affect access to private tutoring, educational materials, and extracurricular opportunities that enrich a student's learning experience.

Motivational And Social Factors

Peer Influence (F score: 178) and Motivation Level (F score: 173): Social dynamics and personal drive both have a noticeable, though moderate, impact on performance. Peer influence can be both positive and negative, depending on the type of social group a student is part of, while motivation plays a critical role in how much effort a student is willing to put into their studies. This suggests that schools might look into creating a positive learning environment and fostering student motivation through rewards or recognition programs.

Less Influential but Notable Factors

1. Teacher Quality (F score: 166): Although we expect teacher quality to have a considerable impact, it ranks lower in importance here compared to factors related to the student's own actions, like studying hours or attendance. This may indicate that while good teachers are important, they can only do so much ultimately, student success largely depends on the effort and support they receive outside of the classroom.
2. Distance from Home (F score: 146): The physical distance from home to school seems to have a minor influence on performance. This might reflect the inconvenience or stress associated with long commutes, which could impact punctuality and readiness for school.
3. Learning Disabilities (F score: 104): Learning disabilities show up as a relatively less important factor here. However, this does not mean they are insignificant, but perhaps the model is not directly targeting interventions for these students. Schools might still focus on providing specialized support to students with learning challenges to help bridge any gaps.
4. Extracurricular Activities (F score: 103): This low score suggests that participation in extracurricular activities does not directly impact academic performance, though these activities can still provide other important benefits, such as social skills, teamwork, and leadership qualities.

Demographic Factors:

1. School Type (F score: 82) and Gender (F score: 74): These demographic factors play a minimal role in determining student performance in this model. This suggests that, regardless of whether a student attends a public or private school, or whether they are male or female, their performance is more strongly shaped by individual actions and support structures.

2. Internet Access (F score: 39): Surprisingly, internet access has the lowest importance, which may indicate that other factors like study habits, parental involvement, or attendance are far more crucial in determining performance than just having access to online resources. However, this finding may vary depending on the context of the educational environment (e.g., rural vs. urban settings).

The Implications Of Feature Importance

The feature importance analysis gives us a wealth of insights into what truly matters when predicting student performance. The dominance of factors like hours studied, previous scores, and attendance underscores the importance of student discipline and consistency in their studies. While personal effort is clearly the biggest driver, factors like sleep, tutoring, and parental involvement demonstrate that a student's overall support system both in and outside of school plays a critical role in shaping their academic success.

This analysis can serve as a valuable tool for educators and policymakers. Schools might focus on improving attendance rates, offering more tutoring programs, and encouraging healthy lifestyles. Parents, too, play a crucial role, and efforts to increase parental engagement and support could have lasting benefits on student outcomes. Additionally, focusing on fostering motivation and providing access to necessary educational resources could help elevate student performance across the board.

By understanding which factors matter the most, we can create targeted interventions that help improve student success and overall educational outcomes.

Actual-Values Vs Predicted-Values Graph

Below is a scatter plot that compares the actual values (on the x-axis) with the predicted values (on the y-axis) from your XGBoost model for student performance prediction. The scatter plot comparing the actual and predicted values plays a crucial role in understanding how well the XGBoost model performs in predicting student performance. By looking at how closely the predicted values align with the actual ones, we can get a sense of the model's strengths, identify areas for improvement, and better evaluate its overall effectiveness. It provides insight into both the accuracy of predictions and the distribution of errors, allowing you to make informed decisions about potential improvements.

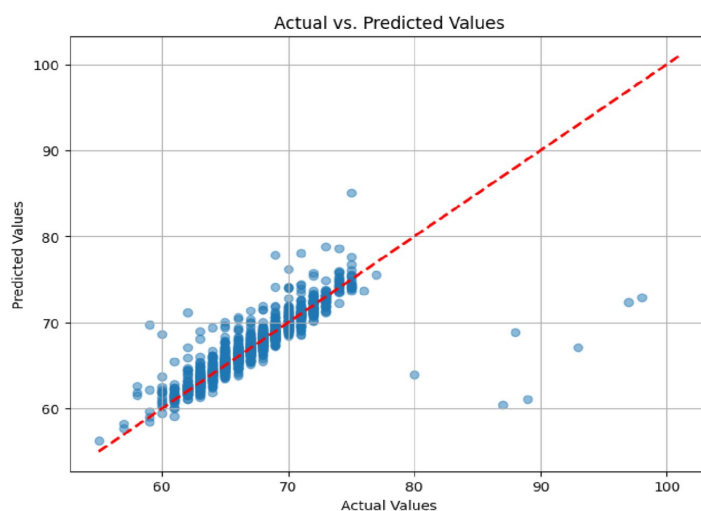


Figure 2: Actual Vs Predicted Value Graph

Key Elements of the Graph

1. **Scatter Points (Blue Circles):** Each blue dot represents an observation, where the x-coordinate is the actual student performance, and the y-coordinate is the corresponding predicted value by the model. Most of the points are concentrated along a diagonal line, indicating that the predictions closely match the actual values, especially in the lower range of scores (around 60-80).
2. **Diagonal Red Line:** The red dashed line represents the ideal scenario where the predicted values are exactly equal to the actual values. If all the points were perfectly aligned along this line, it would indicate a perfect prediction by the model.
3. **Performance Observations:** There is a clustering of points along the diagonal, particularly in the lower range of performance (60 to 80). This suggests that the model has performed reasonably well in predicting these values. However, there are some points that deviate from the red line, especially in the higher range (above 80). These outliers indicate that the model has more difficulty accurately predicting very high actual values. Some predictions overshoot or undershoot these high scores.
4. **Spread and Residuals:** The vertical distance of the points from the red line reflects the residuals, or the difference between the actual and predicted values. While many points are close to the red line (small residuals), there are some outliers that show larger discrepancies between actual and predicted values.

Accuracy

The model shows strong predictive ability for students whose scores fall between 60 and 80. This is evident from the dense cluster of data points near the red diagonal line, which represents a perfect match between actual and predicted values. In this range, the model does a good job capturing the factors that influence student performance, producing accurate predictions for the majority of students. This accuracy suggests that the model has effectively learned patterns from the data such as students' academic histories or engagement in class that are driving their performance within this average to above-average range.

Error Analysis

However, the model seems to struggle with predicting scores for students who perform exceptionally well, particularly those scoring above 80. The scatter plot reveals a few clear outliers in this higher range, indicating that the model's predictions either overestimate or underestimate these students' actual performance. These errors suggest that the model may not be picking up on certain nuances that set top-performing students apart. Factors like study habits, extracurricular involvement, or unique personal circumstances might not be adequately captured in the data the model was trained on. This misalignment with high performers highlights the need for further investigation into what might be causing these discrepancies.

IMPLICATIONS OF FINDINGS

The findings of this study carry significant implications for educators, policymakers, and the broader educational community. For educators, the strong influence of study habits and parental involvement on student performance highlights the need for fostering an environment that supports these factors. Teachers should not only focus on classroom learning but also

encourage students to develop consistent study routines at home. This could involve teaching students time management skills, providing guidance on effective study techniques, and promoting parental engagement. Schools can also step in to support students who may lack access to resources, such as reliable internet, by offering supplementary materials or arranging for community-based support systems.

For policymakers, the results emphasize the importance of investing in digital infrastructure to ensure that all students, regardless of their socioeconomic background, have equitable access to online educational content. In the digital age, internet access is not a luxury but a necessity, and students without it are at a disadvantage. By improving access to technology, policymakers can help bridge the digital divide, providing opportunities for more students to excel academically. Additionally, policymakers should consider implementing programs that encourage greater parental involvement, recognizing that a supportive home environment plays a critical role in student achievement. Community outreach programs, parent-teacher engagement initiatives, and resources for families to better support their children's education could help improve student outcomes across diverse populations.

Challenges and Limitations

During the analysis, several challenges were encountered:

1. **Data Quality:** Some missing values in critical fields (e.g., parental involvement) could have impacted the model's accuracy. Imputing missing values may not have fully captured the relationships between certain features (Amjad et al., 2022).
2. **Imbalanced Classes:** The dataset had an uneven distribution of high, average, and low performers, which could lead to bias in predictions. The model showed a tendency to misclassify at the extremes (high and low performers), which may be due to the small sample size in these categories.
3. **Overfitting Risk:** Given the model's complexity, there was a slight risk of overfitting, particularly when dealing with a large number of features. Further regularization or simplifying the model could mitigate this risk.

Recommendations

Given the model's challenges in predicting scores for high-achieving students, there are several ways its performance could be improved. One option would be to fine-tune the model's settings through hyperparameter optimization to help it make better predictions for outlier cases (Asselman et al., 2023). Another approach could involve introducing additional features into the model. For example, including more detailed data on students' behavior like participation in class, study time, or personal motivation could give the model a better understanding of what drives success. Similarly, considering factors like course difficulty or variations in grading policies might provide more context for why some students perform exceptionally well, helping the model make more accurate predictions for these cases. Additionally, exploring other modeling techniques or combining models through ensemble learning might also reduce prediction errors, particularly for these high performers. Furthermore, as new data becomes available, it is essential to regularly retrain and evaluate the XGBoost model to ensure its predictions remain accurate and relevant. Continuous improvement will enhance the reliability of the insights derived from the model.

CONCLUSION AND FUTURE WORK

The XGBoost model demonstrated reasonable accuracy in predicting student performance, with key features like Study Time, Score, and Attendance being the most influential factors.

The model performed well in identifying average performers but struggled with students on the extreme ends of performance. With further refinement, this model could serve as a useful tool for educators and policymakers looking to improve student outcomes through targeted interventions. For future research, the following points have to be considered in order to improve the model:

1. Explore Alternative Algorithms: Other algorithms such as Random Forest or Neural Networks could be applied to see if they perform better on this dataset. Exploring the integration of other machine learning algorithms, such as deep learning or ensemble methods beyond XGBoost, could provide further insights and improve predictive accuracy.
2. Larger Dataset: A larger dataset with more balanced class distributions (more high and low performers) could improve the model's generalizability.
3. Additional Features: Future studies could incorporate psychological and behavioral features, such as motivation or learning preferences, which might offer deeper insights into student performance.

REFERENCES

- Ahammad, K., Chakraborty, P., Akter, E., Fomey, U. H., & Rahman, S. (2021). A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. *International Journal of Computer Science and Information Security*, 19(1), 5–10. <https://doi.org/10.5281/zenodo.4533374>
- Ajibade, S.-S. M., Ahmad, N. B. B., & Shamsuddin, S. M. (2019). Educational data mining: Enhancement of student performance model using ensemble methods. *IOP Conference Series: Materials Science and Engineering*, 551(1), 012061. <https://doi.org/10.1088/1757-899X/551/1/012061>
- Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., & Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation. *Applied Sciences*, 12(4), 2126.
- Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, 31(6), 3360–3379. <https://doi.org/10.1080/10494820.2021.1928235>
- Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk: Predicting student dropouts using administrative student data and machine learning methods. *SSRN*. <https://doi.org/10.2139/ssrn.3275433>
- Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (Vol. 12163, pp. 129–140). Springer. https://doi.org/10.1007/978-3-030-52237-7_11
- Evangelista, E. D. L., & Sy, B. D. (2022). An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods. *International*

-
- Journal of Electrical and Computer Engineering, 12(5), 5226–5235. <https://doi.org/10.11591/ijece.v12i5.pp5226-5235>
- Gohiya, H., Lohiya, H., & Patidar, K. (2018). A survey of XGBoost system. *International Journal of Advanced Technology and Engineering Research*, 8, 25–30.
- Lone, A. H., & Wani, M. A. (2022). Predicting dropout and academic performance: A boosting ensemble approach.
- Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. M. Ramalho Correia (Eds.), *Trends and applications in information systems and technologies* (Vol. 1365, pp. 166–175). Springer. https://doi.org/10.1007/978-3-030-72657-7_16
- Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, 18, 14. <https://doi.org/10.5334/dsj-2019-014>
- Singh, R., & Pal, S. (2020). Machine learning algorithms and ensemble technique to improve prediction of students' performance. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(3), 3970–3976. <https://doi.org/10.30534/ijatcse/2020/221932020>
- Sonparote, Y. K., Bamnote, G. R., & Ahmed, S. V. (2023). Comparative analysis of various techniques used for predicting student performance. *International Journal of Advanced Research in Science, Communication and Technology*, 173–180. <https://doi.org/10.48175/IJARSCT-7863>
- Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting student's performance using machine learning algorithms. *International Journal of Intelligent Systems and Applications*, 11(12), 34–45. <https://doi.org/10.5815/ijisa.2019.12.04>
- Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes*, 51(2), 916–934. <https://doi.org/10.1108/K-12-2020-0865>