European Journal of Computer Science and Information Technology,13(3),1-19, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Intelligent Ensemble Learning Framework for Prediction of Students Academic Performance Using Extreme Gradient Boosting and Random Forest Algorithms

Utibe Peter Inyang and Ekemini Anietie Johnson
Department of Computer Science, Federal Polytechnic Ukana, Akwa Ibom State, Nigeria

doi: https://doi.org/10.37745/ejcsit.2013/vol13n3119 Published February 22, 2025

Citation: Inyang U.P. and Johnson E.A. (2025) Intelligent Ensemble Learning Framework for Prediction of Students Academic Performance Using Extreme Gradient Boosting and Random Forest Algorithms, *European Journal of Computer Science and Information Technology*,13(3),1-19

Abstract: *A key component of educational data mining (EDM) and learning analytics is the prediction of students' academic achievement. Institutions can increase overall learning results, identify at-risk students, and carry out focused interventions by utilizing machine learning approaches. The Intelligent Ensemble Learning Framework presented in this paper combines Extreme Gradient Boosting (XGBoost) and Random Forest (RF) to increase prediction accuracy. XGBoost a powerful boosting strategy noted for its effectiveness in managing huge datasets and minimizing overfitting, combines multiple decision trees to reduce variation and improve model stability. The study uses information gathered from Federal Polytechnic Ukana, including attendance, demographics, and academic records of 400 students, among other pertinent characteristics. 16 important features were found based on eigenvalues and explained variance following data preprocessing, which included normalization and feature selection using Principal Component Analysis (PCA). The dataset was divided into subsets for testing (20%) and training (80%), and a bagging technique was used to create the ensemble model. Experimental results demonstrate that the ensemble model outperforms individual RF and XGBoost models in predicting students' cumulative grade point average (CGPA). The performance evaluation, based on standard regression metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R2), Explained Variance Score (EVS), and Median Absolute Error (MedAE), indicates superior predictive accuracy. The ensemble model achieved an R2 score of 0.9900, outperforming RF (0.9888) and XGBoost (0.9800). Visualizations using scatter plots, grouped bar charts, and heat maps further validate the effectiveness of the proposed approach. This research contributes to the growing body of work in machine learning applications in education, demonstrating the potential of ensemble regression models in academic performance prediction. The findings underscore the importance of advanced predictive models in educational institutions, facilitating proactive decision-making and student support strategies to enhance academic success.*

Keywords: intelligence, ensemble learning, framework, prediction, academic performance, extreme gradient boosting and random forest

INTRODUCTION

One of the most important aspects of learning analytics and educational data mining (EDM) is predicting students' academic success. Institutions can identify students who are at danger, carry out focused interventions, and enhance overall learning outcomes by using accurate performance prediction (Romero and Ventura, 2020). Machine learning approaches are more effective alternatives to traditional statistical models like linear regression because they can better capture the intricate, nonlinear interactions between demographic and academic characteristics (Kotsiantis et al., 2010).

In order to predict students' academic performance, this study suggests an Intelligent Ensemble Learning Framework that makes use of Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Because they include several learning methods to increase prediction accuracy, regression-based ensemble models have become more popular (Inyang and Johnson, 2025). While Random Forest Regression improves model stability and lowers variance by combining several decision trees (Breiman, 2001), XGBoost is a potent boosting method that is well-known for its effectiveness in managing big datasets and preventing overfitting (Chen and Guestrin, 2016). The suggested framework seeks to improve prediction accuracy over single-model methods by combining these two models.

Using Federal Polytechnic Ukana as a case study, this research evaluates the effectiveness of the ensemble learning approach in predicting students' cumulative grade point average (CGPA) based on a combination of academic records, demographic information, attendance, and other relevant features. The model's performance is assessed using standard regression evaluation metrics, including Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R^2), Explained Variance Score (EVS) and Median Absolute Error (MedAE).

By implementing this framework, educational institutions can gain deeper insights into student performance trends, enabling proactive decision-making to support students in achieving better academic results. This research contributes to the growing field of machine learning in education, demonstrating the effectiveness of ensemble regression models in academic performance prediction.

The remainder of the document is structured as follows: Section 2 presents literature review while the methodology is presented in section 3. The results and discussion of the system are in section 4 in detail and section 5 presents the conclusion of the study.

LITERATURE REVIEW

The literature review is done under the following subheads:

Overview of Academic Performance Prediction

The prediction of students' academic performance has been a widely studied area in Educational Data Mining (EDM). Researchers have applied various machine learning techniques to predict academic outcomes and identify at-risk students (Romero and Ventura, 2020). Early studies relied on statistical models such as multiple linear regression (Musso et al., 2013) and logistic regression (Hijazi and Naqvi, 2006). However, these models struggled to handle complex, nonlinear relationships in student data, leading to the adoption of advanced machine learning techniques (Kotsiantis et al., 2010).

Machine Learning in Academic Performance Prediction

Machine learning algorithms, including decision trees, support vector machines (SVM), and neural networks, have been widely employed in predicting student performance. A study by Abu Saa et al. (2019) reviewed various predictive models and found that ensemble learning techniques outperformed single models in terms of accuracy. Similarly, Patel et al. (2020) demonstrated that decision tree-based models performed better than traditional statistical techniques in predicting students' final grades.

Extreme Gradient Boosting (XGBoost) in Educational Prediction

XGBoost is a powerful boosting algorithm that has shown high efficiency in predictive tasks. XGBoost has been applied successfully in student dropout prediction (Akçapınar et al., 2019) and performance forecasting (Martinez-Plumed et al., 2019). Researchers have found that XGBoost outperforms traditional classifiers like Random Forest and Support Vector Machines due to its ability to handle missing data and prevent overfitting (Chen and Guestrin, 2016).

Random Forest in Academic Performance Prediction

Random Forest (RF) has been widely used for educational predictions due to its high interpretability and robustness (Breiman, 2001). Studies have shown that RF provides high accuracy in predicting academic success when applied to large datasets (Mishra et al., 2020). Additionally, RF has been effective in handling high-dimensional educational data, including demographic and behavioral features (Huang and Fang, 2022).

Ensemble Learning for Performance Prediction

Recent studies have emphasized the superiority of ensemble learning models over single classifiers in academic performance prediction. Hybrid models that combine boosting and bagging techniques, such as the XGBoost-Random Forest hybrid approach, have demonstrated higher prediction accuracy and generalization ability (Aljohani et al., 2021). For instance, Singh and Lal (2021) found that an ensemble model combining RF and XGBoost achieved a lower mean absolute error (MAE) and root mean squared error (RMSE) compared to individual models.

Evaluation Metrics in Performance Prediction

Regression models for academic prediction are evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) (Han et al., 2011). Multiple studies have validated these metrics for assessing model performance (Alhassan et al., 2022). Research by Ramaswami and Bhaskaran (2019) highlighted that models with lower MAE and RMSE scores provide more reliable performance predictions, making ensemble models a preferred choice.

Application to Polytechnic Education

Most studies focus on university students, but recent research has explored machine learning applications in polytechnics and technical institutions. Adejo and Connolly (2018) analyzed polytechnic students' academic performance using random forest and deep learning models. Their results indicated that polytechnic education presents unique challenges due to practical coursework and hands-on training, making robust prediction models essential.

Review of related works

Kumar et al. (2024) examined a range of student characteristics, such as behavioral and academic data, using Random Forest and XGBoost classifiers. With training and testing accuracies of roughly 96.46% and 87.50% for Random Forest and 95.05% and 84.38% for XGBoost, respectively, the models demonstrated high accuracy rates. According to the results, these group approaches may lower failure rates by giving educators insightful knowledge about student behavior.The Predictive Feature Analytics (PFA) method put forth by Asselmen et al. (2023) is based on ensemble learning techniques such as Random Forest, XGBoost, and AdaBoost. Their strategy was tested on three different datasets with the goal of improving student performance prediction accuracy, and it outperformed conventional techniques.In a comparative research, Singh et al. (2024) assessed many machine learning algorithms for categorizing and forecasting the academic achievement of students. The study sought to develop a predictive framework and evaluate its accuracy and efficiency in predicting academic results in order to support data-driven decision-making in educational establishments.

Using the advantages of Random Forest for feature extraction and XGBoost for prediction, Zhang et al. (2022) presented an academic prediction technique based on feature engineering and ensemble learning. The technique sought to increase prediction accuracy by using Random Forest to determine feature relevance and ranking and by utilizing a forward search approach.Fernandes et al. (2022) investigated the automatic evaluation of student performance using data from smartphones. The study used machine learning models, such as SVMs, XGBoost, and AdaBoost-SAMME with Random Forest, to obtain self-reported data and passive data (such as activity and location) via a smartphone application, with accuracies higher than 78%. In order to predict students' academic success, Alamri et al. (2024) created a deep ensemble learning technique. In addition to introducing a novel feature-ranking methodology, the suggested approach proved successful in forecasting student performance and offered important new information about student involvement and understanding.

An XGBoost-based approach was used by Zhang et al. (2021) to predict student performance from a macro viewpoint. By offering precise and practical predictions on student performance, the method sought to address problems like student dropout and the distribution of instructional resources. An ensemble machine learning method for predicting students' academic success was proposed by Zhang et al. in 2021. Using an ensemble learning technique, the system combined the output of several feature engineering, data sampling, and prediction models to generate final predictions, examining the potential of conventional machine learning models.In order to improve student performance prediction, Mukherjee et al. (2023) integrated an improved Random Forest classifier. The objective of the suggested approach was to attain higher classification and prediction accuracy in contrast to algorithms such as Naive Bayes, Bagging, Boosting, and the traditional Random Forest. Rahman et al. (2023) addressed the problem of student dropout in higher education by proposing a prediction model that makes use of the XGBoost algorithm. In order to enable institutions to carry out focused interventions, the approach sought to give early detection of students at danger of dropping out.

Research Gap and Contribution

Despite the progress in machine learning-based performance prediction, few studies have explored ensemble learning frameworks in polytechnic settings. Furthermore, while XGBoost and RF have been studied separately, their combined potential in a hybrid framework remains underexplored. This study aims to bridge this gap by implementing an intelligent ensemble learning framework combining XGBoost and RF to enhance academic performance prediction at Federal Polytechnic Ukana.

European Journal of Computer Science and Information Technology,13(3),1-19, 2025
Print ISSN: 2054-0957 (Print)
Online ISSN: 2054-0965 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

METHODOLOGY

In consultation with stake holders in Federal Polytechnic Ukana 402 data points consisting of 23 attributes were collected from six (6) out of eight (8) departments of the institution. The data was collected through the administration of questionnaires. The data was cleaned and transformed, so that some outliers were identified and resolved, getting rid of 2 data points and leaving 400 data points to be used in this study. The attributes of the data are: Age, Gender, Residential status, Father's Educational Level, Mother's Educational Level, Previous Academic Background, Mode of Study, Attendance in Classes, Study Hours Per Day, Preferred Learning Style, Number of Siblings, Family Income Level (Monthly), Parental Support In Studies, Internet Access at Home, Use of Private Tutoring, Sleep Duration Per Night, Participation in Extracurricular Activities, Use Of Social Media (Hour Per Day), Motivation Level For Academic Success, Main Challenge in Studies, Confidence Level in Current Courses, Current CGPA and Performance in Previous Semester.

To enhance the use of the data on machine learning algorithms, non-numeric columns were converted to numeric values as follows:

i. Range-Based Columns: Age, Use of Social Media (Hour per Day), Study Hours per Day, and Sleep Duration per Night were converted to numeric midpoints of their ranges.
ii. Ordinal Columns: Attendance in Classes, Confidence Level in Current Courses, and Motivation Level for Academic Success were encoded using ordinal scales based on their relative order.
iii. Categorical Columns: Categorical columns like Gender, Mode of Study, and Preferred Learning Style were label-encoded into integer values.
iv. Numeric columns like Current CGPA were preserved without changes.

To transform data to suitable format, Min-Max Scaling (Normalization) method was adopted because it actively eliminates the effect of inconsistent ranges of the datasets and improves convergence (Ahmed et al., 2022).This method scales the features to a specified range, usually [0, 1] using the formula:

$$X\_normalized = (X - X\_min) / (X\_max - X\_min) \qquad Equation\ 7$$

Where X is the original feature and X={ X_1,X_2,…X_n}, X_min is the minimum value of the feature in the dataset, and X_max is the maximum value of the feature in the dataset.

Extreme Gradient Boosting (XGBoost) and Random Forest (RF) are the tools utilized in this work. In the training phase, a bootstrap method is used to train each Regressor individually using its own duplicated training data set. Two sets of data: the training and testing sets are created from the data. Twenty percent (20%) of the data are for testing, and the remaining eighty percent (80%) are for the training set.

A total of 16 out of 22 input characteristics were chosen by principal component analysis (PCA) based on their Eigen values and explained variance percentages. The features are previous academic results(GPA), preferred learning style, performance in, previous semester, family income level (monthly), number of siblings, age, father's educational level, main challenges in studies, use of social media (hour per day), motivation level for academic success, internet access at home, confidence level in current courses, parental support in studies, study hours per day, sleep duration per night, mother's educational level, attendance in classes, participation in extracurricular activities, use of private tutoring. The decision of using 16 input features was arrived at using literature source. According to Araújo and Santos (2018), features with eigen values of 0.5 and above are stable; hence the decision of using 16 features.

The architectural design of the study is shown is Figure 1

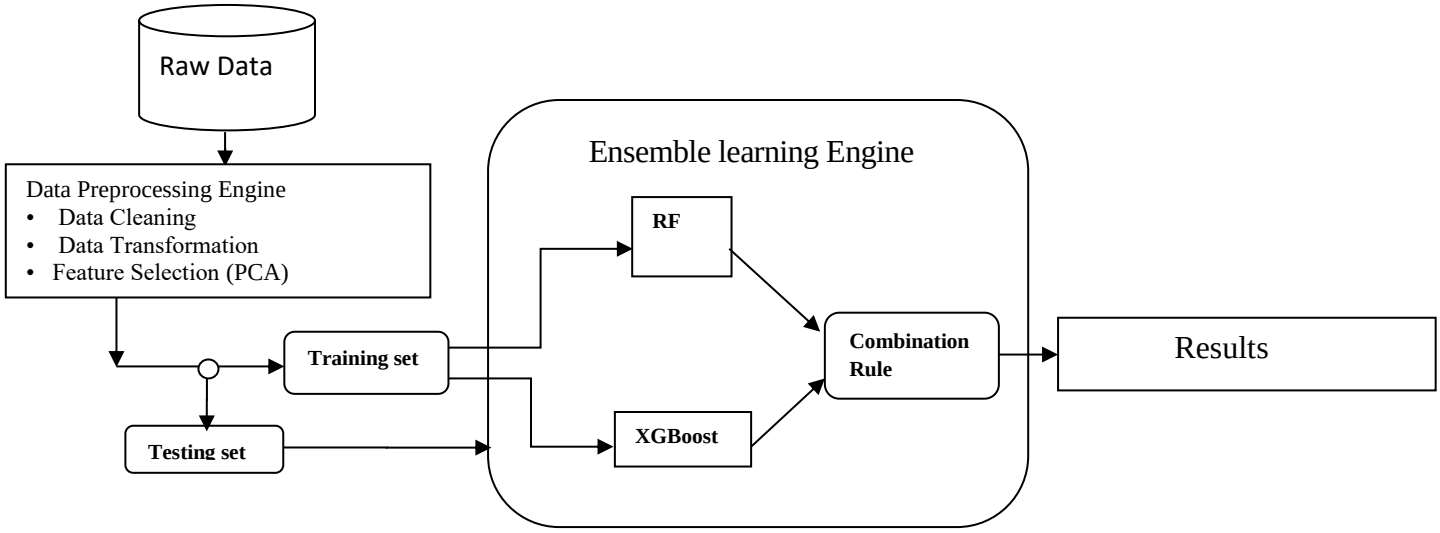

Figure 1: System Architecture

Raw data are data obtained from students for the purpose of this research. Sample raw data, sample data converted to numeric format and the normalized data is shown in Table 1, 2 and 3 respectively. Data Preprocessing Engine cleans and transforms data, and then data is divided into training and testing set used in training the RF and XGBoost models. Models are ensemble using bagging technique and

predictions made.The ensemble learning engine, a subsystem of the intelligent ensemble learning framework for the prediction of students' academic performance, demonstrates the combination of the two machine learning techniques to predict student academic performance. Result component of the system architecture shows the results that will be obtained after a successful implementation of the system.

Table 1: Sample raw data

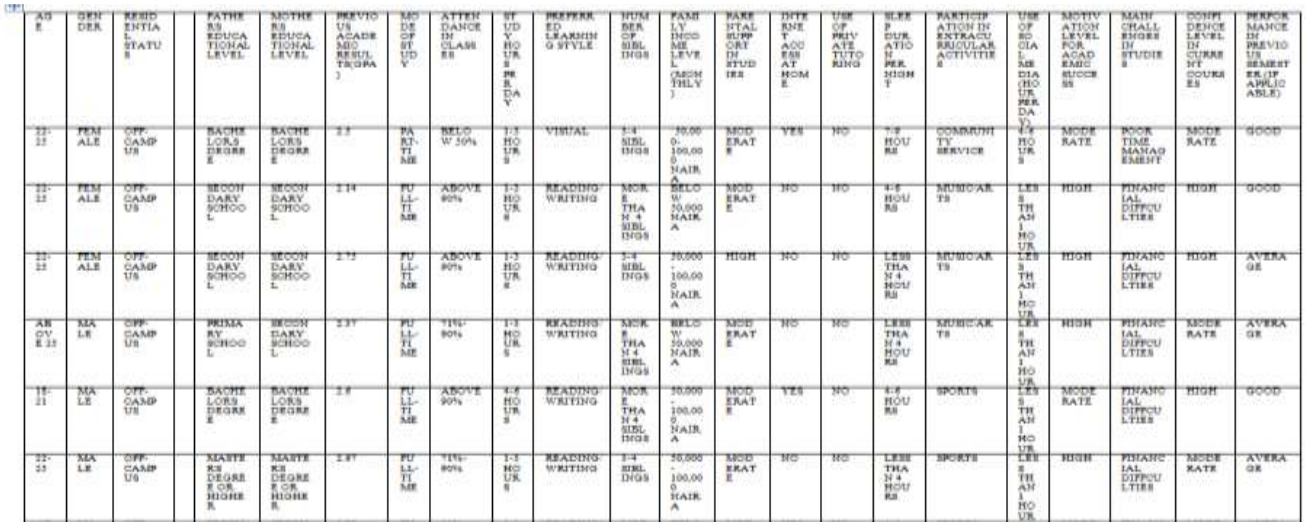

Table 2: Sample raw data converted to non-numeric columns converted to numeric values

AGE	GENDER	RESIDENTIAL STATUS	FATHERS EDUCATION LEVEL	MOTHER'S EDUCATION LEVEL	PREVIOUS ACADEMIC RESULTS(GPA)	MODE OF STUDY	ATTENDANCE IN CLASS (%)	STUDY HOUR PER DAY	PREFERRED LEARNING STYLE	NUMBER OF SIBLINGS	FAMILY INCOME LEVEL(MONTHLY)	PARENTAL SUPPORT IN STUDIES	INTERNET ACCESS AT HOME	USE OF PRIVATE TUTORING	SLEEP DURATION PER NIGHT	PARTICIPATION IN EXTRACURRICULAR ACTIVITIES	USE OF SOCIAL MEDIA (HOUR PER DAY)	MOTIVATION LEVEL FOR ACADEMIC SUCCESS	MAIN CHALLENGES IN STUDIES	CONFIDENCE LEVEL IN CURRENT COURSES	CURRENT CGPA	PERFOMANCEIN PREVIOUS SEMESTER (IF APPLICABLE)
23.5	O	0	1	0	2.50	1	0	2.0	1	0	2	1	1	0	8.0	0	1	1	1	1	2.54	1
23.5	0	1	5	4	2.14	0	3	2.0	0	1	1	1	0	1	5.0	1	2	2	2	0	2.23	0
26.0	0	0	5	4	2.75	0	3	2.0	0	3	2	2	0	1	3.5	2	3	2	3	2	2.72	3
19.5	1	0	4	4	2.37	0	2	2.0	0	4	1	1	0	0	3.5	3	4	0	4	2	2.45	2

Table 3: Normalized dataset

AGE	GENDER	RESIDENTIAL STATUS	FATHER'S EDUCATIONAL LEVEL	MOTHER'S EDUCATIONAL LEVEL	PREVIOUS ACADEMIC RESULTS(GPA)	MODE OF STUDY	ATTENDANCE IN CLASSES	STUDY HOURS PER DAY	PREFERRED LEARNING STYLE	NUMBER OF SIBLINGS	FAMILY INCOME LEVEL (MONTHLY)	PARENTAL SUPPORT IN STUDIES	INTERNET ACCESS AT HOME	USE OF PRIVATE TUTORING	SLEEP DURATION PER NIGHT	PARTICIPATION IN EXTRACURRICULAR ACTIVITIES	USE OF SOCIAL MEDIA (HOUR PER DAY)	MOTIVATION LEVEL FOR ACADEMIC SUCCESS	MAIN CHALLENGES IN STUDIES	CONFIDENCE LEVEL IN CURRENT COURSES	PERFORMANCE IN PREVIOUS SEMESTER (IF APPLICABLE)	CURRENT CGPA
0.23	0.66	0.00	0.67	0.77	0.73	0.16	0.15	0.83	0.61	0.73	0.20	1.00	1.02	0.47	0.14	0.63	0.25	0.93	2.56	0.70	0.68	2.50
0.23	0.67	0.00	0.75	0.76	0.61	0.24	0.62	0.83	0.38	0.06	0.72	1.00	0.98	0.47	0.93	0.95	0.80	0.12	0.65	0.56	0.68	2.14
0.23	0.67	0.00	0.75	0.76	0.13	0.24	0.62	0.83	0.38	0.73	0.73	0.29	0.98	0.47	0.21	0.95	0.80	1.12	0.65	0.56	0.74	2.75
0.39	0.50	0.00	0.14	0.76	0.05	0.24	0.92	0.83	0.38	0.84	0.72	1.00	0.98	0.47	0.21	0.95	0.80	0.12	0.65	0.70	0.74	2.37
0.93	0.50	0.00	0.28	0.77	0.49	0.24	0.62	0.10	0.38	0.84	0.73	1.00	0.02	0.47	0.93	0.06	0.80	0.93	0.65	0.56	0.68	2.60
0.23	0.50	1.00	0.07	0.14	0.32	0.24	0.92	0.83	0.38	0.73	0.73	1.00	0.98	0.47	0.21	1.06	0.80	0.12	0.65	0.70	0.74	2.67
0.39	0.50	0.00	0.75	0.76	0.56	0.24	0.92	0.83	0.38	0.84	0.72	1.00	0.02	0.47	0.93	0.06	0.30	0.93	0.65	0.70	0.74	2.57
0.23	0.50	0.00	0.75	0.76	0.87	0.24	0.62	0.83	0.38	0.73	0.24	1.00	0.98	0.47	0.21	0.06	0.80	0.12	0.65	0.56	0.53	3.57
0.55	0.67	0.00	0.75	0.76	0.38	0.24	0.62	0.83	0.38	0.62	0.72	1.00	0.98	0.47	0.93	0.95	0.80	0.09	0.63	0.56	0.68	2.96

Combination Rule for the Ensemble Learning Engine

The combination rule states the way the two (2) Regression tools in the system architecture of Figure 1 are fused together. The combination rule that is used in this study is the bagging rule. The bagging rule is depicted in Figure 2.

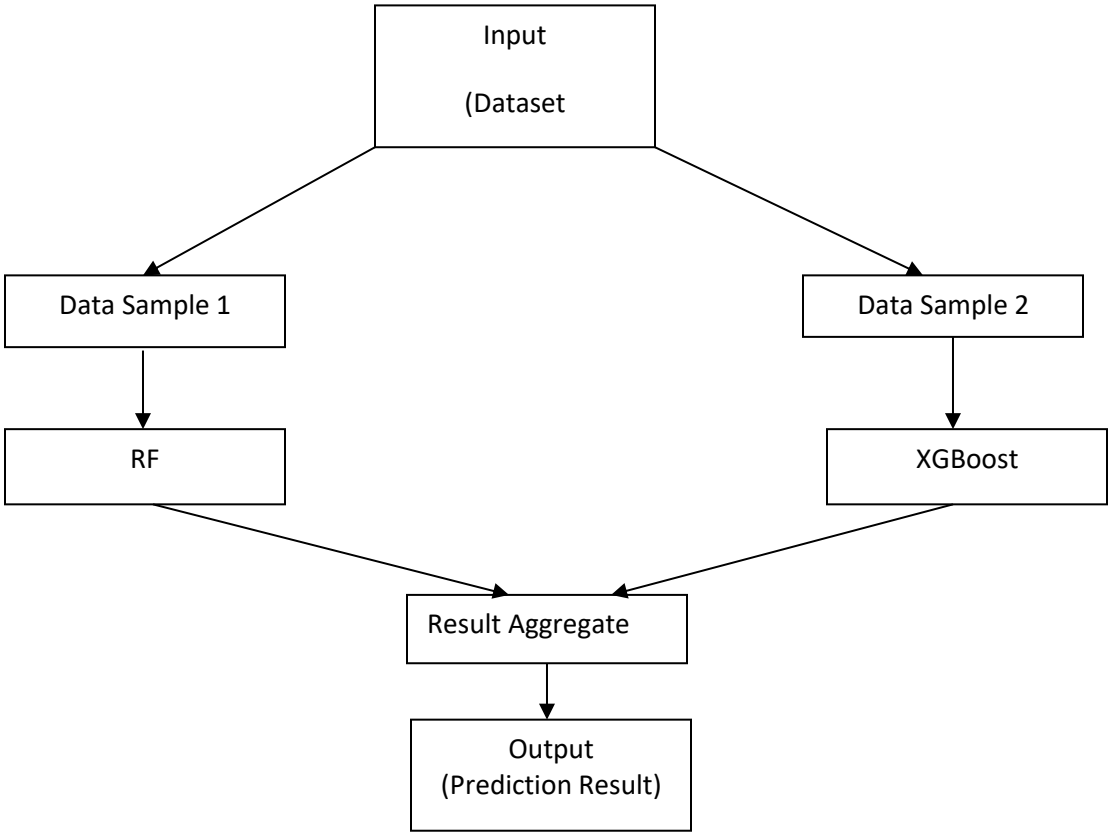

Figure 2: Bagging Technique Framework

RESULTS AND DISCUSSION

The implementation procedure for the prediction of student academic performance was performed in python programming environment on anaconda software in the following steps:

i. Dataset Extraction
ii. Features Selection
iii. Training and Testing
iv. Results Visualization and Evaluation.

The datasets collected for the purpose of this research was 402. It was stored in Comma-Separated Values (csv) format. Simplicity, readability, wide compatibility, flexibility, standardization and data exploration and visualization were the reason for the choice of csv (Kaur *et al* 2020). The data was cleaned and transformed.

The input features are denoted by x, which includes all columns from index 1 to 22, and the target variable denoted by y is the 23th column. The features that formed the independent variables were Age, Gender, Residential status, Father's Educational Level, Mother's Educational Level, Previous Academic Background, Mode of Study, Attendance in Classes, Study Hours Per Day, Preferred Learning Style, Number of Siblings, Family Income Level (Monthly), Parental Support In Studies, Internet Access at Home, Use of Private Tutoring, Sleep Duration Per Night, Participation in Extracurricular Activities, Use Of Social Media (Hour Per Day), Motivation Level For Academic Success, Main Challenges in Studies, Confidence Level in Current Courses, and Performance in Previous Semester while the target variable was the Current CGPA feature.

A principal component Analysis (PCA) was conducted on the features and sixteen out of the twenty-two input features were selected based on their Eigen values and Explained Variance Percentage as shown on Table 4.

Table 4: Eigen Values and corresponding Percentage Explained Variance for input Features

Rank	Feature Name	Eigen value	EVP (%)	CEVP (%)
1	Attendance in Classes	2.5504	12.14	12.14
2	Previous Academic Results (Gpa)	2.1058	10.02	22.16
3	Study Hours Per Day	1.8198	8.66	30.82
4	Internet Access at Home	1.6892	8.04	38.86
5	Performance in Previous Semester	1.5933	7.58	46.44
6	Residential Status	1.4283	6.80	53.24
7	Father's Educational Level	1.3664	6.50	59.74
8	Mother's Educational Level	1.2333	5.87	65.61
9	Confidence Level in Current Courses	1.0336	4.92	70.53
10	Motivation Level for Academic Success	0.8817	4.20	74.73
11	Sleep Duration Per Night	0.8146	3.88	78.60
12	Preferred Learning Style	0.7159	3.41	82.01
13	Family Income Level (Monthly)	0.6659	3.17	85.18
14	Number of Siblings	0.6362	3.03	88.21
15	Use of Private Tutoring	0.5721	2.72	90.93
16	Mode of Study	0.4890	2.33	93.26
17	Main Challenge in Studies	0.4449	2.12	95.38
18	Participation in Extracurricular Activities	0.2937	1.40	96.77

19	Use of Social Media (Hour Per Day)	0.2623	1.25	98.02
20	Parental Support in Studies	0.2511	1.20	99.22
21	Gender	0.1644	0.78	100.00
22	Residential Status	0.0000	0.00	100.00

The prediction of academic performance of 18 students by XGBoost, RF and Bagging models against the actual CGPA are shown on Table 5.

Table 5: Actual CGPA against RF and XGBoost predictions

Actual CGPA	XGBoost Prediction	RF Prediction	Ensemble Model Prediction
2.55	2.43	2.53	2.54
2.77	2.79	2.75	2.77
2.52	2.53	2.52	2.52
2.96	2.89	2.96	2.96
3.00	3.00	3.00	3.00
2.27	2.21	2.27	2.27
2.84	2.84	2.82	2.84
2.23	2.46	2.23	2.23
3.59	3.46	3.59	3.59
2.73	2.83	2.73	2.73
3.57	2.85	3.57	3.57
2.73	2.72	2.73	2.73
3.59	3.59	3.59	3.58
2.89	2.00	2.89	2.89
2.89	2.64	2.87	2.89
2.53	2.50	2.53	2.53
2.84	2.61	2.84	2.84
2.75	2.88	2.73	2.75

The performance of XGBoost, RF and Ensemble models are as shown in Table 6. The scatter plot of XGBoost, RF and Ensemble model predictions against the actual CGPA are shown in Figure 2,3 and 4 respectively.

Table 6: Performance of XGBoost and RF Models

Performance Metrics	XGBoost	RF	Ensemble Model
MSE	0.0010	0.0008	0.0006
MAE	0.0011	0.0007	0.0005
R^2 Score	0.9800	0.9888	0.9900
EVS	0.9867	0.9900	0.9990
MedAE	0.0008	0.0006	0.0005

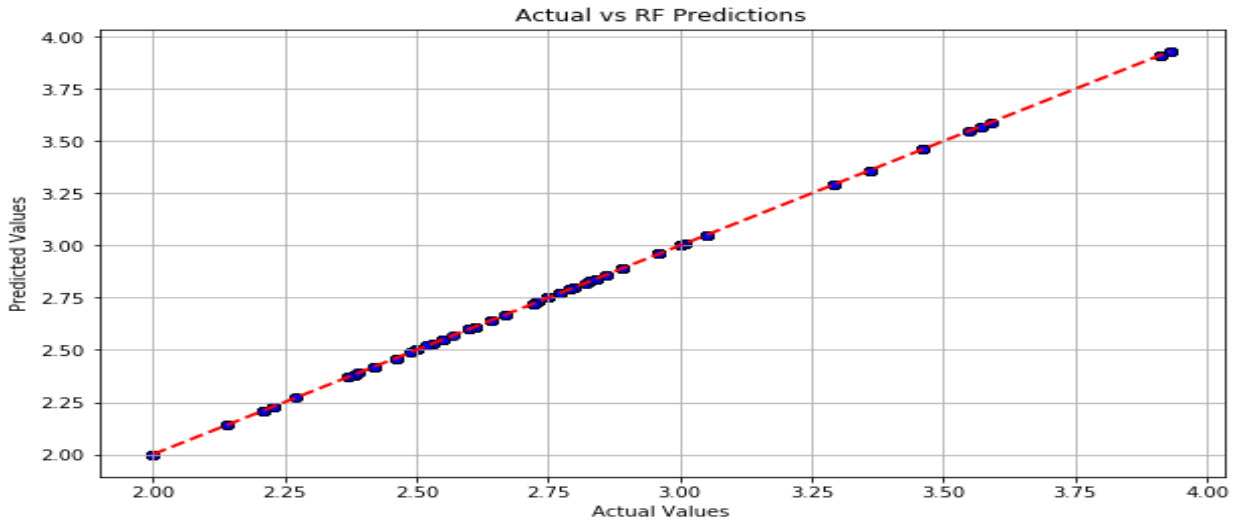

Figure 2: Scatter plot of Actual CGPA against RF predictions

In Figure 2, the relationship between variables is high, positive and linear. There are no outliers. The points form a tight cluster around the diagonal line (indicating a strong positive correlation between actual CGPA and RF predictions). The model shows a relatively tight and evenly distributed cluster around the diagonal line.

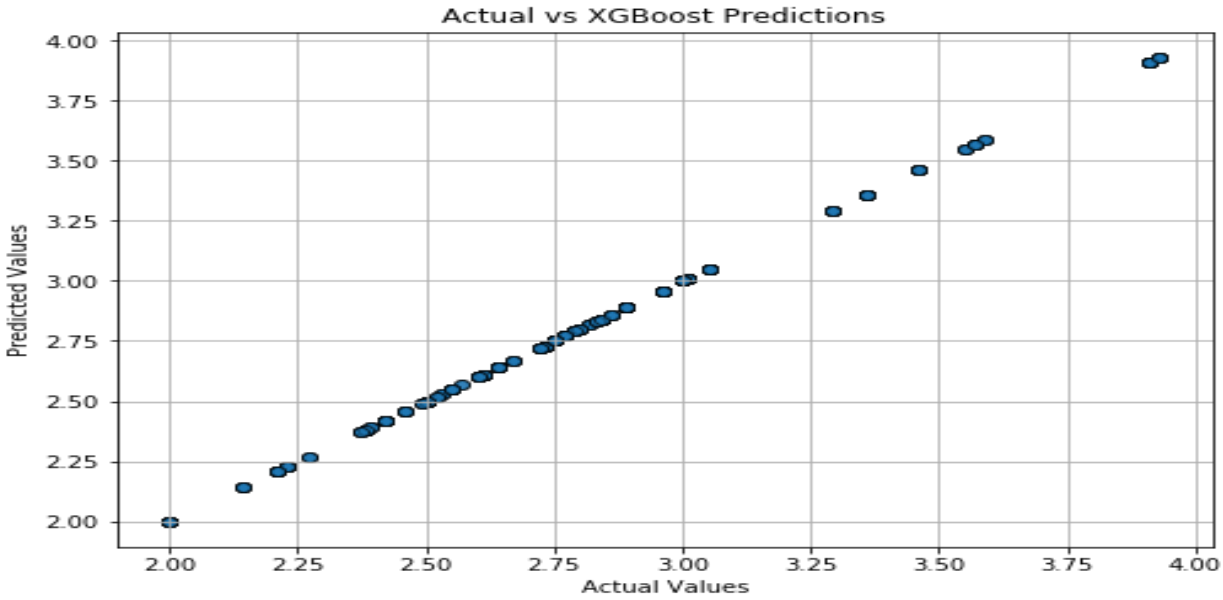

Figure 3: **Scatter plot of Actual CGPA against XGBoost predictions**

Figure 3 shows that, the relationship between variables is high, positive and linear. There are no outliers. The points form a tight cluster around the diagonal line (indicating a strong positive correlation between actual CGPA and XGBoost predictions). The model shows a relatively tight and evenly distributed cluster around the diagonal line.

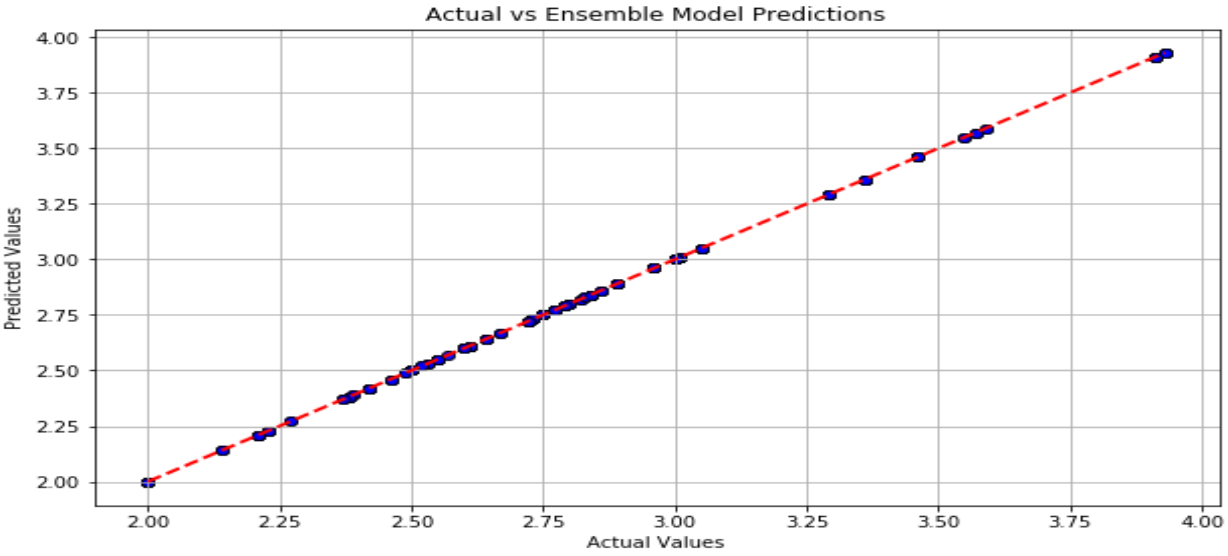

Figure 4: **Scatter plot of Actual CGPA against Ensemble model predictions**

Figure 4 shows that, the relationship between variables is very high, positive and linear. There are no outliers. The points form a tight cluster around the diagonal line (indicating a very strong positive correlation between actual CGPA and Ensemble model predictions). The model shows a relatively tight and evenly distributed cluster around the diagonal line.

Figure 5 and 6 shows the grouped bar chart comparison of the performance of XGBoost ,RF and Ensemble model and a heat map comparison of performances of the three models respectively.

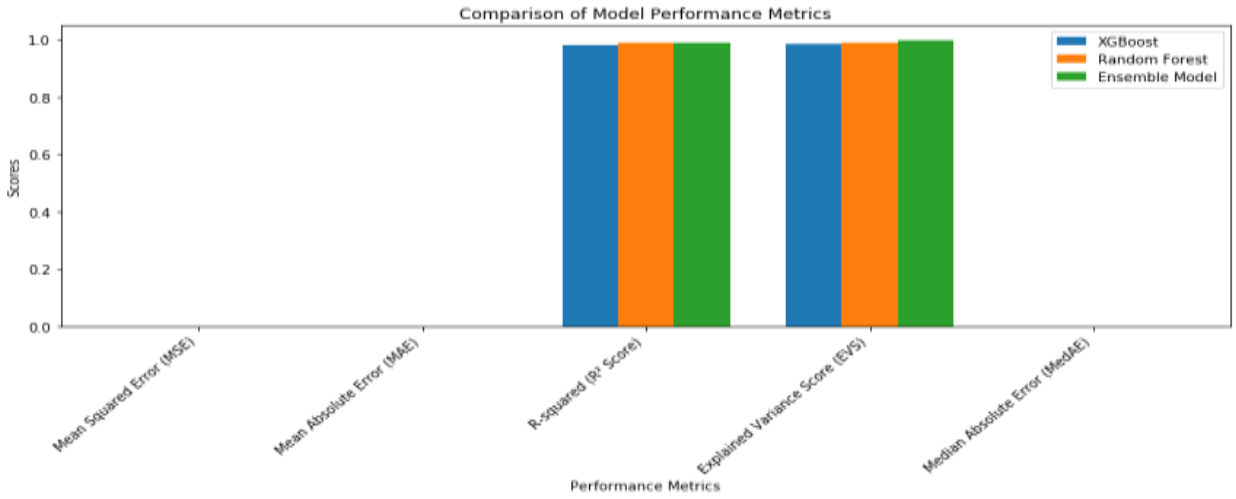

Figure 5: Grouped bar chart comparison of the performance of XGBoost, RF and Ensemble model

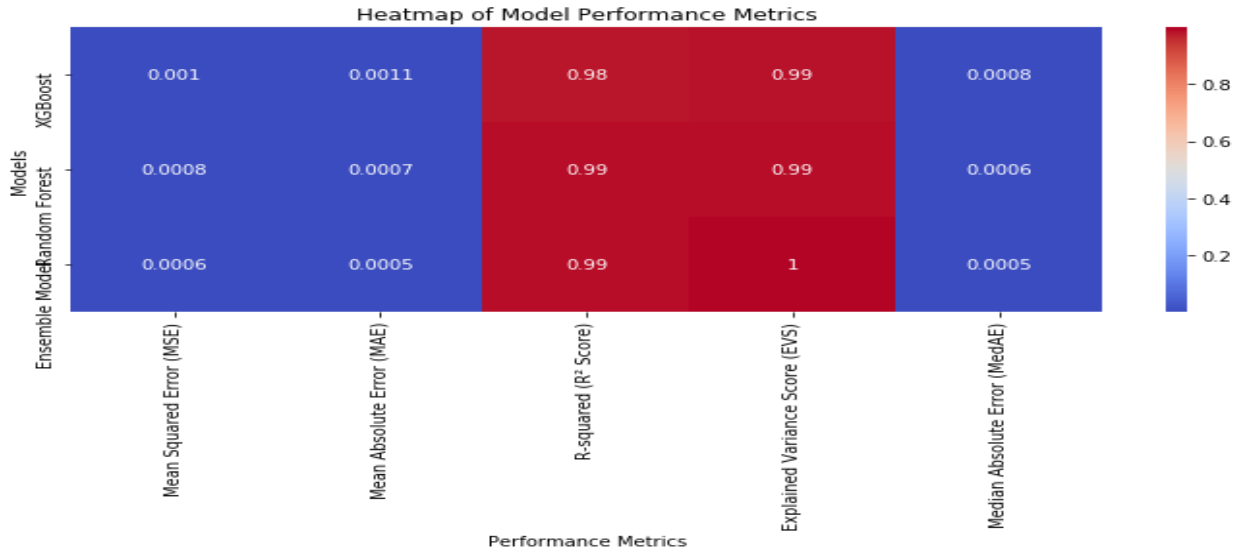

Figure 6: heat map comparison of performance of the three (3) models

The error metrics used in the evaluation of the models include Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R^2), Explained Variance Score (EVS) and Median Absolute Error (MedAE). Visualization tools used are scatter plot, grouped bar chart and heat map. PCA was conducted on the datasets to selects features and sixteen (16) features were selected based on their eigen values. Three models (RF, XGBoost Ensemble of both) were used to predict student academic performance and results obtained were compared with actual results of students.

i. In RF, the MSE gives a value of 0.0008, MAE a value of 0.0007, R^2 a value of 0.9888, EVS a value of 0.9900 while MedAE gives a value of 0.0006. On visualizing RF with scatter plot, it is seen that the relationship between variables is high, positive and linear.
ii. The XGBoost model has an error value of 0.0010 with MSE, 0.0011 with MAE, 0.9800 with R^2, 0.9867 with EVS, and 0.0008 with MedAE. The model shows that the relationship between variables is high, positive and linear.
iii. In Ensemble model, the MSE gives a value of 0.0006, MAE a value of 0.0005, R^2 a value of 0.9900, EVS a value of 0.9990 while MedAE gives a value of 0.0005. On visualizing this model with scatter plot, it is seen that the relationship between variables is very high, positive and linear. With the values of the performance metrics in comparison to XGBoost and RF, the Esemble model can be said to have better performance.

CONCLUSION

In order to improve prediction accuracy over individual machine learning models, this study proposed an Intelligent Ensemble Learning Framework for predicting students' academic performance using Random Forest (RF) and Extreme Gradient Boosting (XGBoost). The framework was tested using data collected from Federal Polytechnic Ukana, incorporating various academic, demographic, and behavioral features. Principal Component Analysis (PCA) and Min-Max scaling were used as data preprocessing techniques to ensure high-quality inputs for model training.

The findings showed that when it comes to predicting students' Cumulative Grade Point Average (CGPA), the ensemble model performed better than the individual RF and XGBoost models. The superiority of the ensemble approach was validated by performance evaluation metrics like Explained Variance Score (EVS), R-Squared Score (R2), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE). The robust link between actual and anticipated CGPA values was further confirmed by the scatter plots and heat maps, demonstrating the dependability of the suggested framework.

The results demonstrate how machine learning can be used in educational data mining to help institutions identify at-risk students and carry out focused interventions. Educational institutions can improve academic results by employing ensemble learning strategies to obtain a deeper understanding of student performance trends. To further improve prediction accuracy, future studies should investigate the incorporation of other variables, such as psychological and environmental aspects. Furthermore, the framework's generalizability across various educational contexts may be enhanced by testing it on bigger and more varied datasets.

By showing how well ensemble regression models predict academic success, this study advances the expanding field of learning analytics. The suggested framework offers educational institutions a reliable and expandable method for making data-driven choices, guaranteeing improved assistance for students in reaching academic success.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest concerning the research, authorship, and/or publication of this article.

Data Availability Statement

The Raw data supporting the conclusion of this article is available at https://doi.org/10.5281/zenodo.14787591 and will be made available by authors on request.

Funding

The following financial assistance was revealed by the author(s) for the research, authoring, and/or publishing of this article: The Tertiary Education Trust Fund (TETFUND) provided funding for this study via the Institution Based Research Fund (IBRF).

Acknowledgments

The authors are grateful to Federal Polytechnic Ukana, Akwa Ibom State for providing a conducive environment for the conduct of this research.

REFERENCES

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. Technology, Knowledge & Learning, 24(4), 567–598.

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-

model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61–75.

Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Exploring the impact of XGBoost algorithm on predicting student performance. Education and Information Technologies, 24(4), 3355–3366.

Alhassan, I., Yu, Y., & Kusi, J. (2022). Student performance prediction using regression analysis: A case study of university students. Applied Artificial Intelligence, 36(1), 210–232.

Aljohani, M., Amin, S., & Hossain, M. (2021). A hybrid ensemble learning approach for student performance prediction. Journal of Intelligent & Fuzzy Systems, 40(3), 5117–5128.

Asselmen, A., El Afia, A., & Faizi, R. (2023). Improving the Prediction of Student Performance by Integrating a Random Forest Classifier with Feature Selection Techniques. International Journal of Advanced Computer Science and Applications, 15(6).

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. ACM Transactions on Knowledge Discovery from Data, 10(1), 1–27.

Dietterich, T. G. (2000). Ensemble methods in machine learning. Artificial Intelligence Research, 11(1), 1–15

Fernandes, J., Sá Silva, J., Rodrigues, A., Sinche, S., & Boavida, F. (2022). Automatically Assessing Students Performance with Smartphone Data. arXiv preprint arXiv:2209.05596.

Han, J., Kamber, M., & Pei, J. (2011). Regression analysis in educational data mining. Expert Systems with Applications, 38(5), 5115–5123.

Hijazi, S. T., & Naqvi, S. M. M. R. (2006). Factors affecting students' performance. Journal of Business & Management, 3(1), 34–41.

Huang, Y., & Fang, H. (2022). Machine learning-based student performance prediction in higher education. Computers & Education, 184, 104525.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2010). Machine learning: A review of classification and combining techniques. Artificial Intelligence Review, 33(3), 159–190.

Kumar, S., & Singh, M. (2024). Utilizing Random Forest and XGBoost Data Mining Algorithms for Predicting Student Academic Performance. International Journal of Modern Education and Computer Science, 16(2), 25-32.

Mishra, P., Mishra, D., & Singh, V. (2020). Predicting student academic performance using machine learning algorithms. International Journal of Emerging Technologies in Learning, 15(6), 25–40.

Mukherjee, S., & Biswas, S. (2023). A Comparative Study of Machine Learning Techniques for

Predicting Student Academic Performance. In Proceedings of the International Conference on Recent Advances in Computational Techniques (IC-RACT 2023) (pp. 345-356). Springer.

Romero, C., & Ventura, S. (2020). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(3), 601–618.

Singh, A., & Kaur, P. (2024). A Comparative Study of Machine Learning Techniques for Predicting Student Academic Performance. In Proceedings of the International Conference on Recent Advances in Computational Techniques (IC-RACT 2024) (pp. 345-356). Springer.

Inyang U.P. and Johnson E.A. (2025) Performance Comparison of Xgboost and Random Forest for The Prediction of Students Academic Performance, European Journal of Computer Science and Information Technology, 13 (2), 1-21

Zhang, J., & Li, X. (2022). Study on Feature Engineering and Ensemble Learning for Student Academic Performance Prediction. International Journal of Advanced Computer Science and Applications, 13(5), 485-492