European Journal of Computer Science and Information Technology, 13 (2), 1-21, 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

Performance Comparison of Xgboost and Random Forest for The Prediction of Students Academic Performance

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/vol13n2121 Published February 11, 2025

Citation: 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

Abstract: *In educational data mining and learning analytics, predicting student academic performance is essential because it provides stakeholders with useful information to improve educational outcomes. In order to predict students' academic results, this study assesses and contrasts the effectiveness of two popular machine learning algorithms: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Data preparation methods, such as principal component analysis (PCA) and feature normalization, were used to enhance a real-world dataset of 400 records gathered from six departments at Federal Polytechnic Ukana. Based on their Eigen values and explained variance, sixteen crucial input features were chosen for examination. Eighty percent (80%) of the dataset was used for training, and the remaining twenty percent (20%) was used for testing. To evaluate the performance of the models, evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R^2), Explained Variance Score (EVS), and Median Absolute Error (MedAE) were used. The findings show that both models have strong predictive powers, with RF marginally outperforming XGBoost in important parameters. The results highlight the potential of data-driven tactics to enhance student outcomes and offer evidence-based suggestions for choosing machine learning models in educational predictive analytics.*

Keywords: performance, extreme gradient boosting, random forest, prediction, students, academic performance.

INTRODUCTION

Predicting student academic performance is a critical task in educational data mining and learning analytics. Accurate predictions can provide valuable insights for educators, administrators, and policymakers to design targeted interventions and improve overall educational outcomes. With the

increasing availability of educational datasets, machine learning techniques have emerged as powerful tools for modeling and predicting academic performance (Kotsiantis et al., 2004).

Two popular machine learning algorithms frequently used in predictive analytics are Extreme Gradient Boosting (XGBoost) and Random Forest (RF). XGBoost, an efficient and scalable implementation of gradient-boosted decision trees, has gained prominence due to its robust performance in competitions such as Kaggle (Chen and Guestrin, 2016). On the other hand, Random Forest, an ensemble learning technique that combines multiple decision trees, has long been a preferred choice for its simplicity and effectiveness in handling high-dimensional data (Breiman, 2001).

The application of these algorithms in educational settings has been explored in various studies. For instance, Sébastien et al. (2018) demonstrated the potential of XGBoost in predicting student success in massive open online courses (MOOCs), highlighting its superior performance compared to other algorithms. Similarly, a study by Saini and Goel (2020) used Random Forest to identify key factors influencing students' academic performance, showcasing its ability to provide interpretable results.

Despite the individual strengths of XGBoost and Random Forest, limited research has directly compared their performance in the context of predicting students' academic outcomes. This study aims to bridge this gap by evaluating and comparing the performance of XGBoost and Random Forest using a real-world educational dataset. The comparison will be based on critical evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-Squared Score (R^2), Explained Variance Score (EVS) and Median Absolute Error (MedAE), offering insights into their applicability and effectiveness in the educational domain.

The findings from this study will contribute to the growing body of literature on educational data mining by providing evidence-based recommendations for selecting appropriate machine learning models in academic performance prediction tasks. Moreover, it will empower stakeholders to make data-driven decisions aimed at enhancing the learning experiences and outcomes of students.

The remainder of the document is structured as follows: Section 2 presents reviewed related literature 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.

European Journal of Computer Science and Information Technology, 13 (2), 1-21, 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

REVIEW OF RELATED LITERATURE

The review of recent works explores existing research and findings related to the prediction of students' academic performance. The prediction of student academic performance has garnered significant attention in educational research, driven by the need to enhance learning outcomes and identify at-risk students early. This literature review explores the current state of research on predictive modeling in education, with a focus on the application of XGBoost and RF algorithms. Early studies in academic performance prediction primarily employed traditional statistical methods, such as linear regression and decision trees, to analyze factors influencing student success. XGBoost emerged as a powerful tool due to its simplicity and effectiveness in handling binary classification tasks, such as pass/fail outcomes. Researchers have utilized XGBoost to identify critical predictors, including attendance, prior academic records, socio-economic status, and engagement levels, demonstrating its utility in educational settings (Johnson et al., 2021).

As data availability and computational power have increased, more sophisticated machine learning techniques have been adopted. The RF algorithm, introduced by Breiman (2001), has become particularly popular for its high accuracy and ability to manage large, complex datasets. Studies employing random forest have reported superior performance in predicting academic outcomes compared to traditional methods, highlighting its robustness against overfitting and its capability to capture nonlinear relationships among variables.

Recent literature has seen a comparative analysis of various predictive models to determine the most effective approaches for specific educational contexts. These studies often emphasize the trade-offs between interpretability and predictive power. While XGBoost offers clear insights into the influence of individual predictors, RF provides a more nuanced understanding through its ensemble approach, albeit with less interpretability.

Furthermore, contemporary research has explored the integration of these models with other advanced techniques, such as neural networks and ensemble methods, to enhance prediction accuracy. The incorporation of feature selection methods and the use of balanced datasets are also discussed extensively, addressing common challenges like multicollinearity and class imbalance.

Rodríguez-Hernández et al. (2021) used artificial neural networks in academic performance prediction. The first objective of this study is to test a systematic procedure for implementing artificial neural networks to predict academic performance in higher education. The second

objective is to analyze the importance of several well-known predictors of academic performance in higher education. The sample included 162,030 students of both genders from private and public universities in Colombia. The findings suggest that it is possible to systematically implement artificial neural networks to classify students' academic performance as either high (accuracy of 82%) or low (accuracy of 71%). Artificial neural networks outperform other machine-learning algorithms in evaluation metrics such as the recall and the F1 score. Furthermore, it is found that prior academic achievement, socioeconomic conditions, and high school characteristics are important predictors of students' academic performance in higher education. Finally, this study discusses recommendations for implementing artificial neural networks and several considerations for the analysis of academic performance in higher education.

Johnson et al., (2024) built an intelligent analytic framework for predicting students academic performance using multiple linear regression and random forest. After thorough data preparation and standardization, 664 datasets from eight departments at Federal Polytechnic Ukana were used in the study. Metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared Score (R^2), and Explained Variance Score (EVS) were used to assess the performance of both models. The findings showed that RF performed noticeably better than MLR, with higher predicted accuracy and reduced error rates. Bar charts and scatter plots provided more evidence of RF's strong performance over MLR. This study highlights how incorporating cutting-edge machine learning methods into classroom environments can yield better understandings of student performance and allow for more focused and timely interventions. The results support the use of RF in order to improve educational outcomes and provide more accurate forecasts.

A model for predicting student performance based on supervised machine learning techniques was created by Hashim et al. in 2020. A number of supervised machine learning algorithms, including Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Sequential Minimal Optimization, and Neural Network, were compared in this study. In order to predict student performance on final exams, the researchers trained a model using datasets from courses in the bachelor study programs at the College of Computer Science and Information Technology, University of Basra, for the academic years 2017–2018 and 2018–2019. The best classifier for precisely predicting students' final grades, according to the results, is the logistic regression classifier (68.7% for passed and 88.8% for failed).

Johnson et al. (2023) did a comparison of two machine learning techniques for the prediction of initial oil in place in the Niger Delta region. Shell Petroleum Development Company (SPDC)

provided 816 datasets for this study, which were predicted using the volumetric approach. These data sets were preprocessed and applied to two machine learning methods: support vector regressor and random forestto estimate the initial oil in place. The outcomes were contrasted with those from SPDC. Compared to calculations using all nine features, the findings of calculations utilizing four of the nine major features were more similar to those from SPDC. Additionally, support vector regressor and random forest computation outcomes were evaluated. The random forest findings show a stronger correlation (0.970) with the field results than the support vector regressor (0.832). This study is unique in that it uses four predicting features (independent variables) to provide prediction values that are extremely similar to those found in the field using nine features. Random forest was used to get this, making it a dependable machine technique for predicting the amount of oil that will initially be present in the Niger Delta.

Hashim et al. (2020) predicted students' academic performance using ensemble approaches. We collected educational data from a learning management system (LMS) in order to illustrate the significance of student behavioral aspects in this article. The included dataset was subjected to feature analysis, followed by data preprocessing—a crucial phase in the knowledge-discovery process. The preprocessed dataset is classified using classifiers including Nave Bayes (NB), Decision Tree (ID3), Support Vector Machines (SVM), and K-Nearest Neighbor (KNN) in order to predict student academic achievement. The suggested model's accuracy is increased through the application of ensemble methods. We employed typical ensemble techniques including bagging, boosting, and voting algorithms. By employing group methods, we were able to improve the outcome and show the dependability of the suggested model.

Nabil et al. (2021) predicted students' academic performance based on courses' grades using deep neural networks. The main goal of this paper is to explore the efficiency of deep learning in the field of EDM, especially in predicting students' academic performance, to identify students at risk of failure. A dataset collected from a public 4-year university was used in this study to develop predictive models to predict students' academic performance of upcoming courses given their grades in the previous courses of the first academic year using a deep neural network (DNN), decision tree, random forest, gradient boosting, logistic regression, support vector classifier, and K-nearest neighbor. In addition, we made a comparison between various resampling methods to solve the imbalanced dataset problem, such as SMOTE, ADASYN, ROS, and SMOTE-ENN. From the experimental results, it is observed that the proposed DNN model can predict students' performance in a data structure course and can also identify students at risk of failure at an early

stage of a semester with an accuracy of 89%, which is higher than models like decision tree, logistic regression, support vector classifier, and K-nearest neighbor.

Using artificial neural networks, Lau et al. (2019) predicted and categorized the academic performance of their students. Eleven input variables, two levels of hidden neurons, and one output layer make up the neural network model. The backpropagation training rule is implemented using the Levenberg-Marquardt algorithm. The area under the receiver operating characteristics curve, error performance, regression, error histogram, confusion matrix, and error histogram are used to assess the effectiveness of the neural network model. Despite certain drawbacks, the neural network model has an overall strong prediction accuracy of 84.8%.

In order to predict students' academic performance, Albreiki (2021) conducted a mining of student information system records. The primary goal of this research is to determine which characteristics that influence students' performance are most frequently researched and which data mining approaches are most frequently used to find these factors. As a result, a dataset from a nearby university in the United Arab Emirates' student information system was created for this dissertation. The dataset, which had a record count of over 56,000, had 34 attributes relating to student information. According to empirical findings, four categories of student characteristics such as demographics, past performance history, course and teacher information, and some general student information—are in charge of predicting academic success. Furthermore, the findings also showed that artificial neural networks, decision trees, and Naïve Bayes are the most often utilized data mining methods for categorizing and predicting student variables. The best data-mining model for forecasting students' academic success from student information systems was ultimately determined by comparing a set of models.

Tomasevic, et al (2020) aimed of at providing a comprehensive analysis and comparison of state of the art supervised machine learning techniques applied for solving the task of student exam performance prediction, i.e. discovering students at a “high risk” of dropping out from the course, and predicting their future achievements, such as for instance, the final exam scores. For both classification and regression tasks, the overall highest precision was obtained with artificial neural networks by feeding the student engagement data and past performance data, while the usage of demographic data did not show significant influence on the precision of predictions. To exploit the full potential of the student exam performance prediction, it was concluded that adequate data acquisition functionalities and the student interaction with the learning environment is a prerequisite to ensure sufficient amount of data for analysis.

In summary, the literature reflects a dynamic and evolving field, with ongoing efforts to refine predictive models and adapt them to diverse educational environments. This review aims to synthesize these advancements, providing a comprehensive understanding of the current methodologies and identifying gaps for future research. Through this synthesis, we seek to establish a foundation for our comparative study of multiple XGBoost and Random Forest in predicting student academic performance.

Random Forest

Random Forest is a popular machine learning algorithm and an ensemble learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability (Wainberg et al., 2016). A Random Forest is a tree-based ensemble with each tree depending on a collection of random variables. More formally, for a p-dimensional random vector $X = (X_1, \ldots, Xp)^T$representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution $P_{XY}(X,Y)$. The goal is to find a prediction function $f(X)$ for predicting Y. The prediction function is determined by a loss function $L(Y, f(X))$ and defined to minimize the expected value of the loss.

$$E_{XY}(L(Y, f(X))) \quad \text{Equation 1}$$

where the subscripts denote expectation with respect to the joint distribution of X and Y.

Intuitively, $L(Y, f(X))$ is a measure of how close $f(X)$ is to Y; it penalizes values of $f(X)$ that are a long way from Y. Typical choices of L are *squared error loss* $L(Y, f(X)) = (Y - f(X))^2$ for regression and *zero-one loss* for classification:

$$L(Y, f(X)) = I\big(Y \neq f(X)\big) = \begin{cases} 0 \; if \; Y = f(X) \\ 1 \; otherwise \end{cases} \quad \text{Equation 2}$$

It turns out that minimizing E_{XY} (L (Y, f (X))) for squared error loss gives the conditional expectation

$$f(x) = E(Y|X=x) \quad \text{Equation 3}$$

Otherwise known as the *regression function*. In the classification situation, if the set of possible values of Y is denoted by Y, minimizing $E_{XY}(L(Y, f(X)))$ for zero-one loss gives:

$$f(x) = argmax P(Y=y|X=x) \quad \text{Equation 4}$$

otherwise known as the *Bayes rule*.

Ensembles construct f in terms of a collection of so-called "base learners" $h_1(x), \ldots, hJ(x)$ and these base learners are combined to give the "ensemble predictor" $f(x)$. In regression, the base learners are averaged

$$f(x) = \frac{1}{J}\sum_{J=1}^{J} h_j(x) \quad \text{Equation 5}$$

$$f(x) arg\max_{y \in Y} \sum_{J=1}^{J} I\,(y = h_j(x)) \qquad \text{Equation 6}$$

Extreme Gradient Boost

The XGBoost (Extreme Gradient Boosting) algorithm is an optimized and scalable implementation of gradient-boosted decision trees (GBDT), which is designed for speed and performance. It has become one of the most popular machine learning algorithms due to its efficiency, flexibility, and ability to handle a variety of data types and problems.

The Working of XGBoost can be explained using the following steps:

a. Initialization: Starts with an initial prediction, usually the mean (for regression) or a uniform distribution (for classification).
b. Gradient Descent Optimization: Each subsequent tree is trained to minimize the residual errors (gradients) of the previous predictions.
c. Tree Construction: Trees are built iteratively, where splits are determined based on the reduction of the loss function (e.g., Mean Squared Error for regression, Logarithmic Loss for classification).
d. Weighted Learning: Each tree assigns weights to instances, giving higher importance to incorrectly predicted examples.
e. Final Prediction: Combines the predictions of all the trees (via weighted sums for regression or probability scores for classification) to make the final output.

XGBoost has the following advantages:

i. High Performance: Delivers state-of-the-art results in competitions and benchmarks.
ii. Flexibility: Works with multiple loss functions and is extensible with user-defined objectives.
iii. Scalability: Handles large datasets and high-dimensional data efficiently.
iv. Robustness: Built-in features like regularization and early stopping help prevent overfitting.

XGBoost ca be applied in Classification problems (e.g., credit risk analysis, fraud detection), Regression problems (e.g., price prediction, sales forecasting), Ranking problems (e.g., search engine result ranking), Time-series forecasting and Feature selection through its feature importance scores.

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 Challenges 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=$\{ X_1, X_2, \ldots X_n \}$, 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

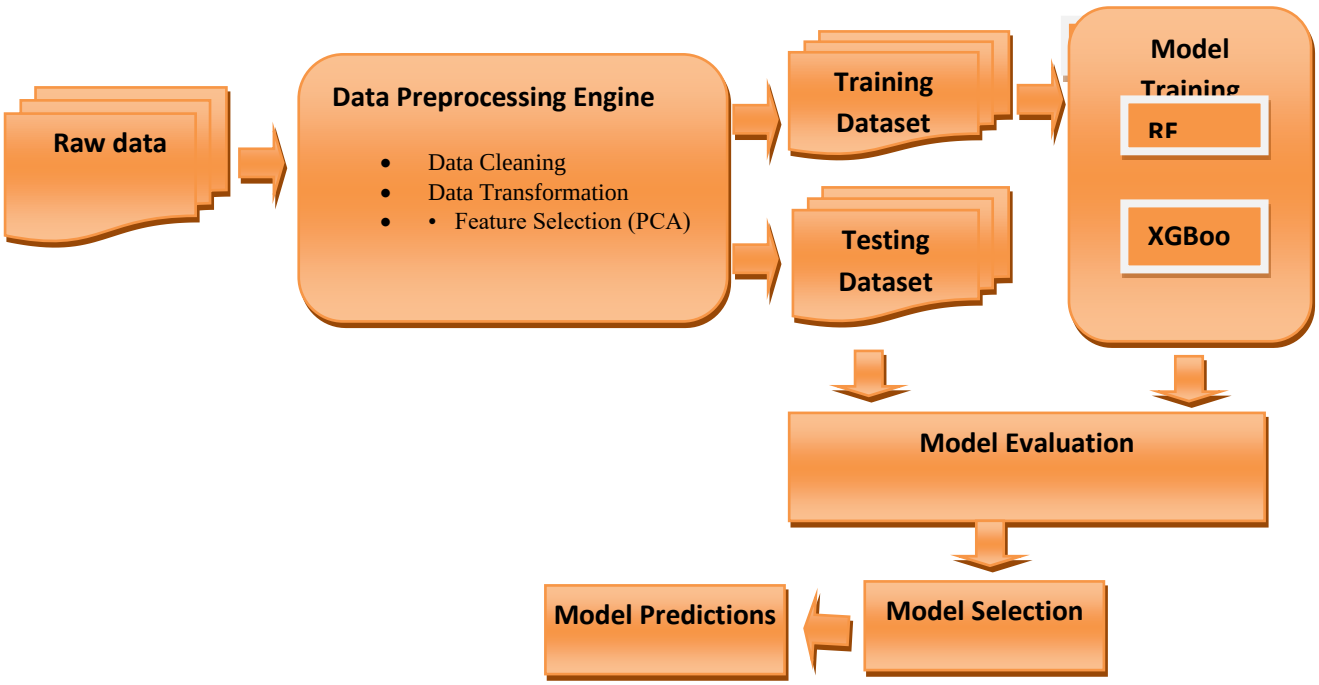

Figure 1.0: System Architecture

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, then data is divided into training and testing set used in training the RF and XGBoost models. Models are evaluated and selected and the model that performs better is used in prediction.

Table 1: Sample raw data

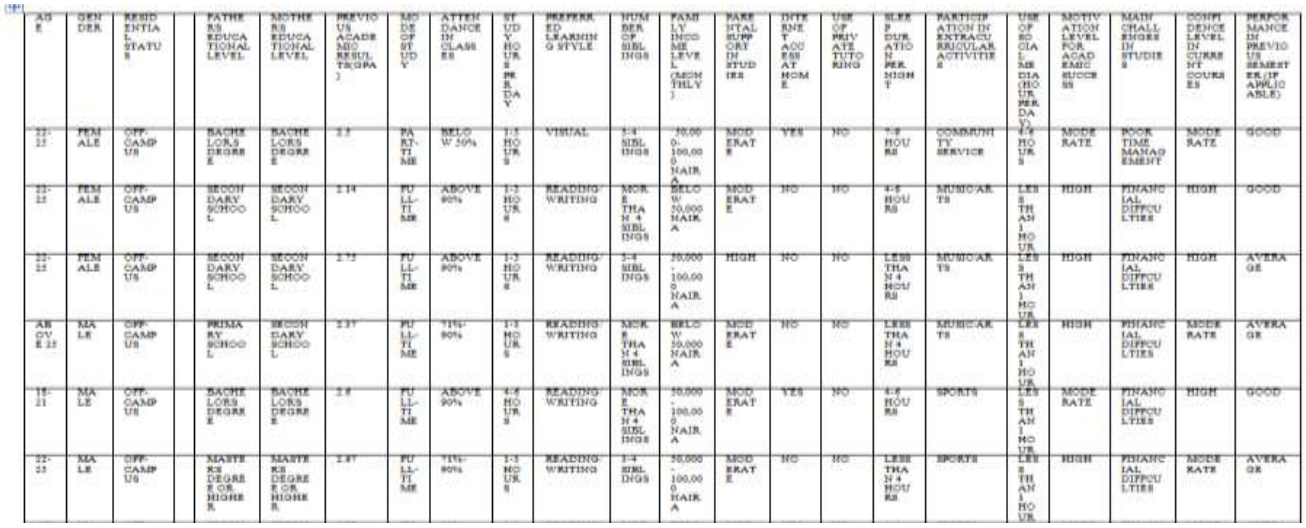

Table 2: Sample raw data converted to non-numeric columns converted to numeric values

AGE	GENDER	RESIDENTIAL STATUS	FATHERS EDUCATION LEVEL	MOTHER'S EDUCATION LEVEL	PREVIOU 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	PERFORMANCEIN 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	5	4	2.75	0	3	2.0	0	3	2	2	0	1	3.5	2	3	2	3	2	2.72	3

.0																						
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

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 from 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 Challenges 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 RF and XGBoost against the actual CGPA are shown on Table 5.

Table 5: Actual CGPA against RF and XGBoost predictions

Actual CGPA	XGBoost Prediction	RF Prediction
2.55	2.43	2.53
2.77	2.79	2.75
2.52	2.53	2.52
2.96	2.89	2.96
3.00	3.00	3.00
2.27	2.21	2.27

2.84	2.84	2.82
2.23	2.46	2.23
3.59	3.46	3.59
2.73	2.83	2.73
3.57	2.85	3.57
2.73	2.72	2.73
3.59	3.59	3.59
2.89	2.00	2.89
2.89	2.64	2.87
2.53	2.50	2.53
2.84	2.61	2.84
2.75	2.88	2.73

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

Table 6: Performance of XGBoost and RF Models

Performance Metrics	**XGBoost**	**RF**
Mean Squared Error (MSE)	0.0010	0.0008
Mean Absolute Error (MAE)	0.0011	0.0007
R-squared (R^2 Score)	0.9800	0.9888
Explained Variance Score (EVS)	0.9867	0.9900
Median Absolute Error (MedAE)	0.0008	0.0006

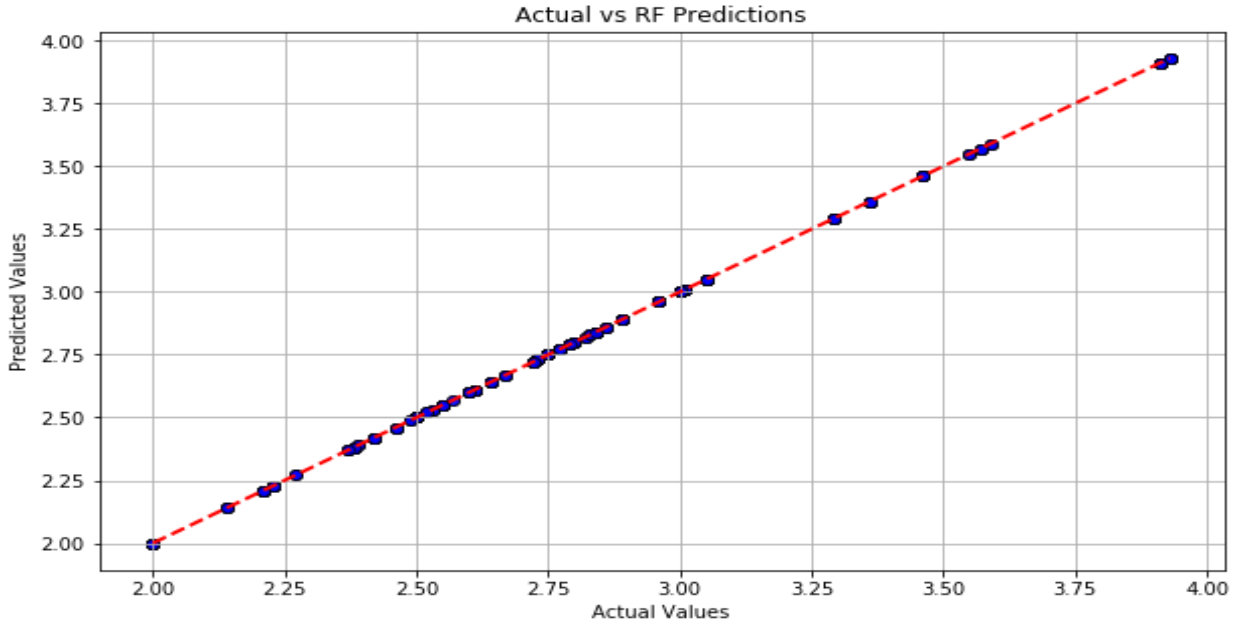

Figure 2: Scatter plot of Actual CGPA against RF predictions

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.

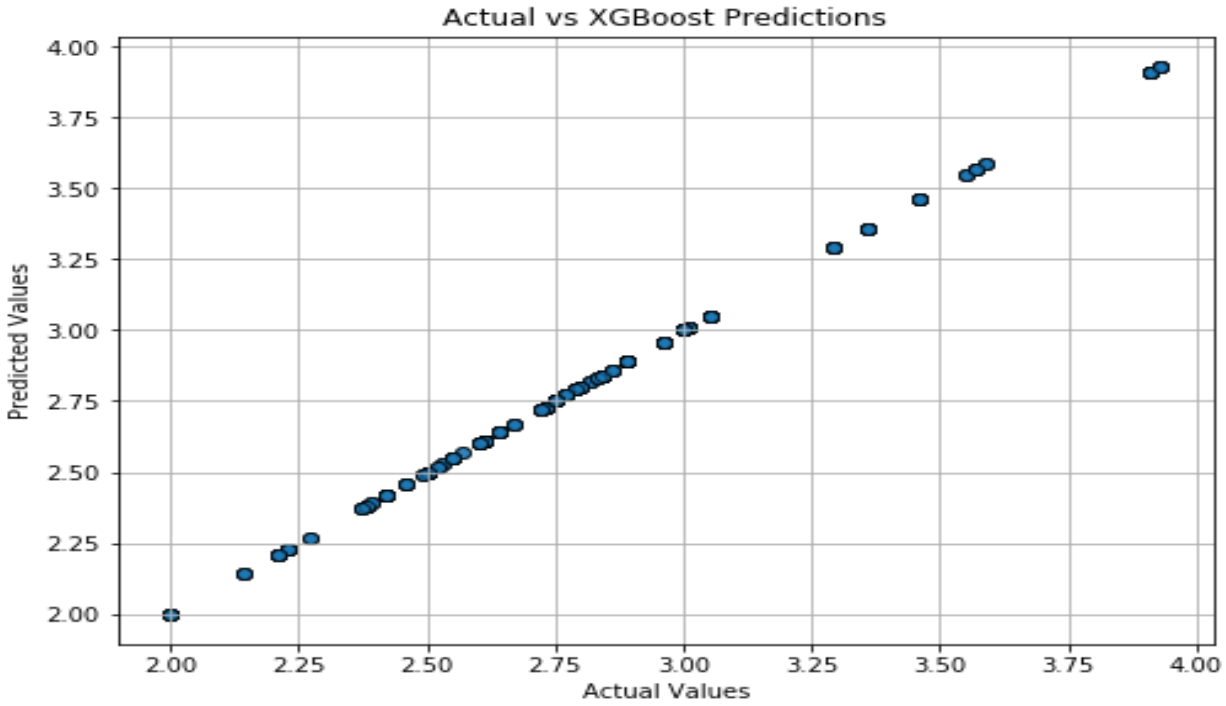

Figure 3: **Scatter plot of Actual CGPA against XGBoost predictions**

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 shows the grouped bar chart comparison of the performance of XGBoost and RF.

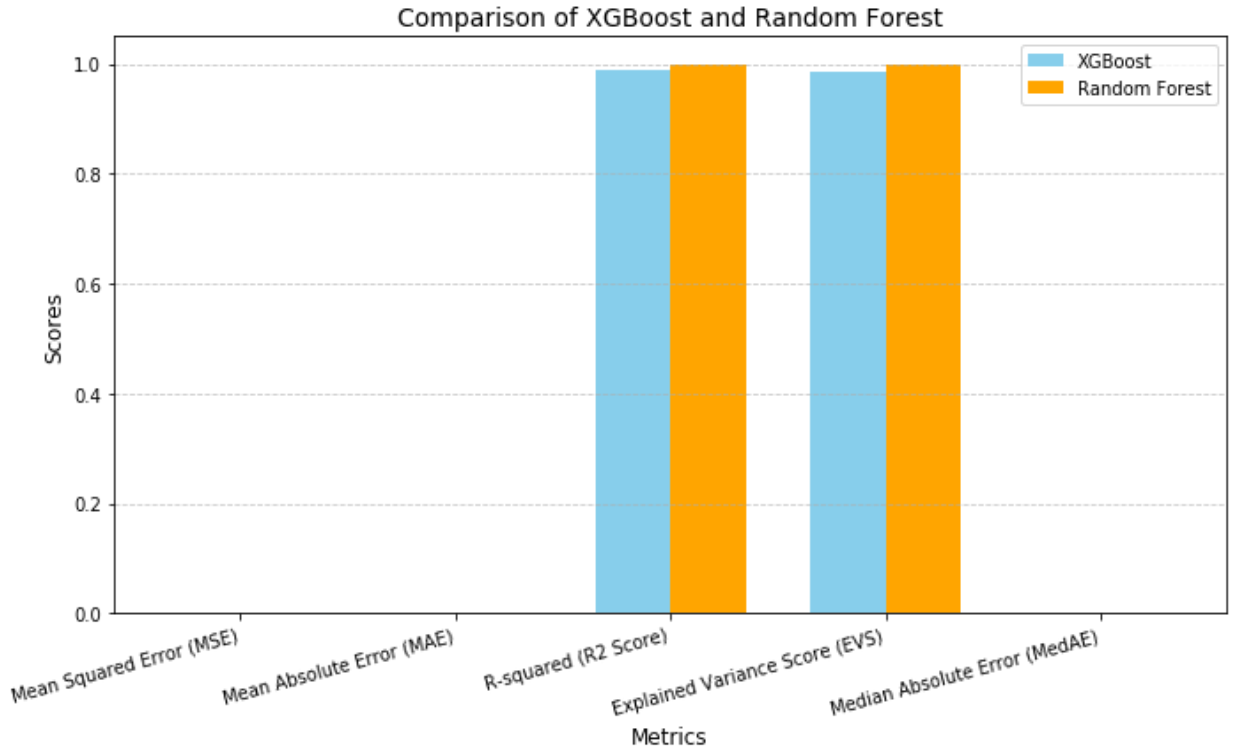

Figure 4: Grouped bar chart showing performance of XGBoost and RF.

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 and grouped bar chart. PCA was conducted on the datasets to selects features and sixteen (16) features were selected based on their eigen values. Two models (RF and XGBoost) 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. With the values of the performance metrics in comparison to XGBoost, RF model can be said to have better performance.

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. With the values of the

performance metrics in comparison to RF, XGBoost model can be said to have a good performance.

CONCLUSION

This study demonstrates the utility of machine learning algorithms, specifically XGBoost and Random Forest, in predicting student academic performance using a comprehensive dataset. Both models exhibited strong performance, with RF achieving marginally better results across most evaluation metrics. The principal component analysis (PCA)-driven feature selection process proved effective in identifying the most influential predictors, emphasizing the importance of data preprocessing in achieving high model accuracy.

Given the findings, the following recommendations are proposed:

i. Adopt RF for predictive analytics: Due to its superior performance and scalability, RF is recommended as the primary algorithm for academic performance prediction tasks, especially in scenarios with large datasets and complex relationships.
ii. Invest in data-driven decision-making: Educational institutions should prioritize collecting and maintaining high-quality, diverse datasets to leverage advanced machine learning techniques effectively.
iii. Expand research scope: Future studies should explore hybrid modeling approaches that combine the strengths of XGBoost and Random Forest to further enhance predictive accuracy.
iv. Integrate predictive insights into academic Policies: Policymakers and educators should utilize model insights to design targeted interventions aimed at improving academic success, focusing on key predictors such as previous academic results, attendance, and study habits.

The study highlights the transformative potential of machine learning in educational settings, paving the way for more personalized and effective academic strategies.

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

Ahmed, H. A., Ali, P. J. M., Faeq, A. K. and Abdullah, S. M. (2022). An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method. *Aro-The Scientific Journal of Koya University*, 10(2): 29-37.

Araújo, J. M. and Santos, T. L. M. (2018). Control of a class of second-order linear vibrating Systemswith time-delay: *Smith predictor approach. Mechanical Systems and Signal Processing*, 108: 173-187.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020). Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.

Johnson, E. A., Inyangetoh, J. A., & Esang, M. O (2021). An Experimental Comparison of Classification Tools for Fake News Detection. International Journal of Advanced Research in Computer and Communication Engineering Vol. 10, Issue 8, August 2021 DOI 10.17148/IJARCCE.2021.10820

Johnson E., Obot O., Inyang U.,and Akpabio J.. (2023) Comparison of Two Machine Learning Techniques for the Prediction of Initial Oil in Place in the Niger Delta Region, European Journal of Computer Science and Information Technology 11 (5), 30-49

Johnson E.A., Inyangetoh J.A., Rahmon H.A., Jimoh T.G., Dan E.E., Esang M.O. (2024) An Intelligent Analytic Framework for Predicting Students Academic Performance Using Multiple Linear Regression and Random Forest, European Journal of Computer Science and Information Technology, 12 (3),56-70

Kaur, A., Ayyagari, S., Mishra, M. and Thukral, R. (2020). A Literature Review on Device-to-Device Data Exchange Formats for IoT Applications.*Journal of Intelligent Systems and Computing*, 1(1): 1-10.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411-426.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1(9), 982.

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9, 140731-140746.

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. Computers and Education: Artificial Intelligence, 2, 100018.

Saini, M., & Goel, S. (2020). Predicting academic performance of students using Random Forest algorithm. *Procedia Computer Science, 167*, 2331-2340.

Sébastien, B., Courtemanche, F., & Lamontagne, L. (2018). Predicting student success in MOOCs using XGBoost. *International Journal of Emerging Technologies in Learning, 13*(12), 101-109.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & education, 143, 103676.

Wainberg, M., Alipanahi, B., and Frey, B. J. (2016). Are random forests truly the best classifiers?. *The Journal of Machine Learning Research*, *17*(1), 3837-3841.

European Journal of Computer Science and Information Technology, 13 (2), 1-21, 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