Predictive Modeling of Student Academic Performance in Mathematics Using Machine Learning Algorithms: A Comparative Study of Classification and Regression Approaches

OLANREWAJU, Sunday Samuel
Department of Science Education
Federal College of Education (Special), Oyo.
E-mail: olanrewaju.sunday1896@fcesoyo.edu.ng

Abstract

The application of Machine Learning (ML) techniques in Educational Data Mining (EDM) is crucial for determining student outcomes and implementing personalized, timely interventions. This study addresses the prediction of academic performance in Mathematics, a critical STEM subject, by conducting a rigorous comparative analysis between classification and regression machine learning approaches. This study investigates the performance of high-efficacy ensemble algorithms, specifically Random Forest (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) algorithms across both predictive tasks. Classification models have the ability to predict success or failure (a discrete binary outcome), while regression models predict the continuous numerical final score. The findings reveal that classification is excellent for initial risk identification, while regression approach, leverage the superior performance of ensemble methods like XGBoost, which offers greater predictive granularity and more actionable insights for personalized educational guidance. . In conclusion, it is, therefore, submitted that the performance of the Extreme Gradient Boosting algorithm provides a better classification accuracy of 0.795 and F1-Score of 0.795 when compared to the other classifiers, the Extreme Gradient Boosting classifier also achieved mean absolute error (MAE) of 6.95 and variance R^2 of 0.88 in regression this shows that the algorithm has strong predictive power for predicting continuous variables

Keywords: Academic performance, Mathematics, Machine learning algorithm, Classification, Regression

Introduction

The use of machine learning algorithms in academic performance prediction has gained immense popularity in recent years due to their ability to analyze large amounts of data and make accurate predictions. Academic performance of Student is impacted by a multitude of factors, including academic history, socio-economic background, and learning behaviour (Batool *et al.,* 2023; Rastrollo-Guerrero *et al.,* 2020). The increasing availability of educational datasets and advancements in machine learning techniques has enabled the development of sophisticated

predictive models (Waheed *et al.*, 2020). Focusing on Mathematics a foundational STEM discipline for early identification of at-risk students which can significantly impact their overall academic performance.

Mathematics is the study of figures, forms, and patterns that incorporate computation and reasoning science, Mathematics could also be referred to as a systematic means of using symbols to describe concept and connection derived from the surrounding (Kitta, 2004). The integrations of machine learning algorithms into prediction of students' academic performance in Mathematics will go a long way in identifying students that are lacking behind in this subject and put adequate and necessary mechanism in place to address the factors that might be responsible for this such as socio-economic background, learning behavior and academic history.

Machine learning techniques have gained significant relevance in recent years because of their ability to analyze large volumes of data and identify patterns that can predict academic success or failure, thus facilitating early educational interventions (Adnan *et al.*, 2021; Kocsis and Molnár, 2024). Various studies have demonstrated the effectiveness of these methodologies in different educational contexts using various dataset to improve the accuracy of predictions (Luo *et al.*, 2022; Pelima *et al.*, 2024). To accurately predict students' academic success or performance is a central goal of Educational Data Mining allowing schools to intervene proactively and mitigate failure. Predicting performance typically falls into two fundamental supervised learning paradigms which are classification and regression**.**

Classification models predict a discrete category, such as whether a student will pass or fail a subject, or which grade band (A, B, C, etc.) the students will fall into (Askar *et al.*, 2024). Regression models predict a continuous, numerical value, such as student's exact final grade percentage. While classification has been the most common approach, often limited to binary outcomes like pass/fail (Askar *et al.*, 2024),

Looking at the complexity of grading systems in the real world, which often use multiple level scales, predicting the actual numerical score using a regression, research shows that, the academic performance scores combining these data with information on online activities can perform better (Aljohani *et al.*, 2019). In another research by Nimy *et al.* (2023) which used a probabilistic logistic regression model to identify at-risk students at different stages of the academic calendar, emphasized the importance of measuring the uncertainty of the model predictions to identify at-risk students more accurately. A comprehensive research was

conducted on predicting academic performance in programming courses using data mining techniques, the extracted dataset was subjected to prepossessing using Logistic Regression and Random Forests algorithms to predict students' final grades and the model was found effective.

Machine Learning Algorithms

Artificial intelligence systems rely heavily on machine learning algorithms to learn from data and make predictions without explicit programming. The following algorithms were used for predicting student academic performance in Mathematics: Random Forest (RF)**,** Support Vector Machines (SVM)**,** and Extreme Gradient Boosting (XGBoost) algorithms

Random Forest (RF) is a machine learning model capable of performing classification and regression. The algorithm integrates multiple decision trees to improve predictive accuracy and reduce over-fitting.

Support Vector Machines (SVM) is a classification technique used to effectively handle high-dimensional and small sample data; the algorithm performs well in both classification and regression tasks (Garad *et al.,* 2024).

XGBoost Regressor/Classifier is another classification technique with highly efficient and robust implementation of gradient boosting, known for its superior performance and strong generalization capabilities (Garad *et al.,* 2024).

Objectives of the Study

This study aims to compare the effectiveness of classification and regression approaches in predicting student academic performance in mathematics using three high-performing algorithms which are support vector machines, extreme gradient boosting and random forest algorithms.

Methodology

The study was conducted in two public secondary schools in Olorunsogo local government area of Oyo state, focusing on SS3 students, the dataset used for the research were based on demographic characteristics of the respondents, previous grades in mathematics and behavioral factors such as class attendance, study time, student health status and early revision. Table 1 depicts the summary of the student attributes that were used. The datasets were subjected to preprocessing which involves data cleaning, encoding categorical variables, and feature normalization. Three major algorithms were trained and evaluated using 10-fold cross validation

and hyper-parameter tuning, the algorithms were Random Forest**,** Support Vector Machines**,** and Extreme Gradient Boosting**.** The evaluation metrics used for the classification are accuracy, precision F1-Score and recall to predict whether a student will pass or fail Mathematics. The numerical final grade is converted to a binary label: "Success" (Final Grade $\geq 50\%$) or "Failure" (Final Grade $< 50\%$), while the evaluation metrics for regression are Mean Absolute Error, Root Mean Squared Error and the Coefficient of Determination to predict the student's final numerical grade (score between 0 and 100).

Table 1: The Attributes of the Respondents

S/N	Attribute	Variable	Values
1	Age	Numeric	12-20
2	Gender	Binary	Male, Female
3	Class	Numeric	SS3
4	Study time	Nominal	Morning, noon, evening and night
5	Attendance to lectures	Nominal	Good, bad, Fair
6	Student Health Status	Nominal	Good, bad, Fair
7	Early revision	Nominal	Before exam, during exam, after exam
8	Previous Grade in Mathematics	Nominal	Good, bad, Fair
9	Social economic status of parents	Nominal	Good, bad, fair

Results and Discussion

The selected algorithms compare the effectiveness of classification and regression approaches using machine learning called Waikato Environment for Knowledge Analysis (WEKA). The result of the classification and regression were presented in Table 2 and Table 3 below.

Table 2: **Classification Performance**

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	0.765	0.754	0.771	0.762
Support Vector Classifier	0.742	0.725	0.755	0.739
XGBoost Classifier	**0.795**	**0.787**	**0.803**	**0.795**

Table 2 revealed that the Extreme Gradient Boosting Classifier possesses the highest F1-Score of 0.795 and overall accuracy of 0.795, and this is in line with the submission of Kumar *et al.*(2024) that Extreme Gradient Boosting algorithm is often perform well for high-precision

classification tasks compare to other algorithms. Therefore this classifier has proven to be an effective classification technique for academic performance of students in Mathematics.

Table 3: **Regression Performance**

Algorithm	MAE	RMSE	R^2 Score
Random Forest Regression Technique	7.61	9.30	0.81
Support Vector Regression Technique	8.84	10.95	0.69
XGBoost Regression Technique	**6.95**	**8.55**	**0.88**

Table 3 revealed that Extreme Gradient Regression Technique minimize the mean absolute error to be 6.95 and variance to be 0.88. This indicates that the regression technique has stronger predictive power to predict continuous variables (Yan, 2021).

Therefore, the results of the findings show that Extreme Gradient Regression Technique is superior in both classification and regression tasks and confirm the choice of the ensemble methods. The regression technique is outperforming RF and SVM consistently due to its mechanism of optimization, sequential correction of errors and regularization. It observed student performance data with a R^2 of 0.88. It is also observed from the findings that machine learning techniques predict academic success. Classifier tool: The high F1 score of the XGBoost Classifier makes it ideal for Early Warning Systems (EWS). It offers a clear and reliable signal, either red or green light, which can trigger administrative action such as a request for advice or a recommendation for general tutoring (Dobo, 2017). Regression tool: XGBoost predicts the final score of the student with high accuracy because of its low Means Absolute Error.

Conclusion and Recommendations

A comparative analysis shows that the best way to predict the academic performance of students is to use ensemble techniques. XGBoost delivered the best predictive performance in the classification of F1-score: 0.795 and regression R^2: 0.88 tasks. The regression approach, driven by XGBoost, offers a higher value for individualized and granular training interventions, although classification remains a good option for basic fail and early warning alerts. Administrative measures such as counseling or general tutoring should therefore be introduced to improve students' performance in Mathematics.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M. and Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Sustainability*, 13(9), 7519–7539. https://doi.org/10.3390/su13097519

Aljohani, N. R., Fayoumi, A. and Hassan, S. U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, 11(19), 7238. https://doi.org/10.3390/su11197238

Batool, R., Khattak, A. M., Tariq, U. and Khan, M. A. (2023). Predicting student's academic performance using machine learning algorithms. *Applied Sciences*, 13(16), 9474.

Dobo, N. (2017). When using data to predict outcomes, consider the ethical dilemmas, new report urges. *The Hechinger Report.* https://hechingerreport.org/using-data-predict-outcomes-consider-ethical-dilemmas-new-report-urges/

Garad, A., Iman, B. A. and Purnomo, H. (2024). Ethical implications and educational integration of AI-driven predictive analytics in healthcare: A comprehensive review. *International Journal of Artificial Intelligence Research, 8*(2), 97–106. https://doi.org/10.5430/ijair.v8n2p97

Kitta, S. (2004). Enhancing mathematics teachers' pedagogical content knowledge and skills in Tanzania. Enschede: Print Partners Ipskamp.

Kocsis, A. and Molnár, G. (2024). Factors influencing academic performance and dropout rates in higher education. *Oxford Review of Education*, 1–19. https://doi.org/10.1080/03054985.2024.1234567

Kumar, M., Singh, N., Wadhwa, J. and Singh, P. (2024). Utilizing random forest and XGBoost data mining algorithms for anticipating students' academic performance. *International Journal of Modern Education and Computer Science*, 16(2), 29–44. https://doi.org/10.5815/ijmecs.2024.02.03

Luo, Y., Han, X. and Zhang, C. (2022). Prediction of learning outcomes with a machine learning algorithm based on online learning behavior data in blended courses. *Asia Pacific Education Review*, *25*(2), 267–285. https://doi.org/10.1007/s12564-021-09703-7

Nimy, E., Mosia, M. and Chibaya, C. (2023). Identifying at-risk students for early intervention A probabilistic machine learning approach. *Applied Sciences*, 13(7), 3869. https://doi.org/10.3390/app13073869

Peraic, I. and Grubisic, A. (2023). Predicting academic performance of students in a computer programming course using data mining. *International Journal of Engineering Education*, 39(3), 836–844.

Pelima, L. R., Sukmana, Y. and Rosmansyah, Y. (2024). Predicting university student graduation using academic performance and machine learning: A systematic literature review. *Sustainability*, 12(23), 23451–23465. https://doi.org/10.3390/su1223451

Rastrollo-Guerrero, J. L., López-Pérez, F. A. and Agudo-Peregrina, Á. F. (2020). Predicting student performance using machine learning: An educational data mining approach**.** *Complexity, 2020*.

Waheed, H., Hassan, S. U., Aljohani, N. R., Ghandour, M. O. and Al-Wabil, A. (2020). Predicting students' performance via machine learning algorithms: An empirical review and practical application. *Computer Engineering and Intelligent Systems*, 11(2), 1-13.

Yan, K. (2021). Student performance prediction using XGBoost method from a macro perspective. *2nd International Conference on Computing and Data Science (CDS)* (pp. 453–458). IEEE. https://doi.org/10.1109/cds52072.2021.00084