Modelling Mathematics Performance of Ghanaian Junior High School Students Using Supervised Machine Learning

Dominic Obeng Koranteng*[1], Mary Akua Agyeiwaa Wiafe[1], and Dr. Bright Emmanuel Owusu[1]

[1]African Institute for Mathematical Sciences (AIMS), Ghana

September 11, 2025

Abstract

Improving student performance in mathematics is a critical goal for educational systems, particularly in developing nations. This study applies supervised machine learning techniques to model and predict the mathematics performance of Junior High School students in Ghana, aiming to identify key predictive factors. We collected data from 364 students using a questionnaire that covered 26 variables related to student demographics, socio-economic background, study habits, and school environment. Four machine learning models were developed: Linear Regression and Random Forest Regressor to predict continuous scores, and Decision Tree and Support Vector Machine (SVM) classifiers to predict performance categories (Pass/Fail). Using a 70/30 train-test split and 5-fold cross-validation, our results show that the Random Forest Regressor significantly outperformed Linear Regression (R^2 = 0.51 vs. 0.16; MAE = 10.71 vs. 16.27). For classification, the Decision Tree model achieved 80% accuracy, proving more effective than the SVM, which struggled with the minority class (failing students). Feature importance analysis revealed that **class size, distance to school, teacher's rating, study hours, and teacher quality** were the most significant predictors of performance. These findings demonstrate the potential of machine learning to provide data-driven insights for targeted educational interventions and policy-making.

Keywords: Machine Learning, Student Performance, Educational Data Mining, Mathematics Education, Predictive Modelling, Ghana

1 Introduction

Mathematics proficiency is a cornerstone of cognitive development, fostering critical analytical and problem-solving skills essential for national progress. However, in many developing countries, including Ghana, students frequently face challenges in mathematics, leading to low achievement rates [1, 3, 24]. Understanding the multifaceted factors that influence academic performance—spanning socio-economic conditions, instructional quality, and student habits—is crucial for developing effective interventions.

The advent of machine learning (ML) offers powerful analytical tools to uncover complex, non-linear patterns in educational data that traditional statistical methods may miss [26, 7]. While research applying ML in education is growing, many studies focus on general academic performance or use limited datasets, leaving a gap in subject-specific, context-aware predictive modelling [2].

*Corresponding author: dobeng@aims.edu.gh

This study addresses this gap by applying supervised machine learning algorithms to a comprehensive dataset from Junior High School students in Ghana. Our primary objectives were:

1. To identify the key factors influencing student performance in mathematics.
2. To evaluate and compare the predictive accuracy of different ML models.
3. To provide actionable, data-driven recommendations for educators and policymakers.

By focusing on a specific subject within a defined regional context, this research contributes a nuanced understanding of performance drivers and offers a robust methodological framework for future educational analytics.

2 Literature Review

The application of machine learning in educational data mining has grown significantly, offering a powerful alternative to traditional statistical methods for predicting academic outcomes. The ability of ML models to handle complex, high-dimensional, and non-linear interactions is particularly valuable in a field where student success is influenced by a wide array of interconnected factors [26]. A comprehensive review by [26] noted that models such as decision trees, support vector machines (SVMs), and artificial neural networks (ANNs) consistently demonstrate superior predictive power over conventional statistical tools. This review also highlighted a growing trend towards using ensemble methods, which combine multiple base models to enhance accuracy and robustness, a key motivation for the use of Random Forest in our study.

Recent empirical studies have reinforced these findings. For instance, [7] utilized Random Forest and Gradient Boosting models to predict academic performance in Pennsylvania's schools, achieving high accuracy. Their work underscored the multifaceted nature of academic achievement, identifying school characteristics, family engagement, and socio-economic status as critical determinants. Similarly, a systematic literature review by [25] argued that prediction models are significantly enhanced when academic data is integrated with behavioural, psychological, and demographic information, thereby creating a holistic and more accurate picture of student outcomes. Further advancing the methodology, [34] introduced a novel graph-based ensemble machine learning approach that achieved 89% accuracy by explicitly modelling the interaction effects between various student, teacher, and environmental parameters. This highlights the importance of not just identifying predictive variables, but also understanding their complex interrelationships.

Research focused specifically on the factors influencing mathematics performance reveals a complex interplay of cognitive, social, and environmental variables. Using data mining techniques on PISA data, [1] identified teacher quality, family support, and the learning environment as universally important factors, which are consistent with the predictors ("Quality of Teaching," "Family's Financial Support," "Classroom Conditions") included in this study. The importance of context was further emphasized by [3] in a comparative analysis of four developing countries. They found that the predictive power of certain factors, such as classroom ventilation or the availability of learning materials, varied significantly between nations. This finding strongly supports the need for localized studies, such as ours in Ghana, that can account for regional and cultural specifics rather than relying on generalized models. Beyond environmental factors, psychological elements also play a crucial role. [24], using structural equation modelling, highlighted the importance of motivation, enjoyment, and academic self-concept—constructs that correlate directly with variables in our dataset like "Do You Enjoy Studying Mathematics (DYESM)" and "Study Hours (SH)."

The choice of ML model is also a critical consideration. An increasing body of research compares the performance of various algorithms in educational contexts. For example,[20], in

a study predicting success in an online mathematics learning game, several algorithms were tested, including Naïve Bayes, Random Forest, and Logistic Regression. They concluded that tree-based models, like Random Forest, generally performed better, particularly when dealing with datasets containing both numerical and categorical features and exhibiting class imbalance—characteristics common to our dataset. While highly predictive, some advanced models like ANNs can be "black boxes." [36] demonstrated the power of ANNs in capturing intricate non-linear relationships in academic and behavioural data, but such models often lack the interpretability needed to provide actionable feedback to educators. This motivated our choice to focus on models like Decision Trees and Random Forest, which provide clear feature importance rankings.

In addition to academic and institutional data, recent studies have increasingly incorporated behavioural and psychosocial predictors. [17] found that student behaviours such as regular attendance and homework completion, along with parental involvement, were more predictive of academic success than prior academic records alone. This aligns with the inclusion of variables like "Do You Miss School (DYMS)" and "How Often Do You Do Your Homework (HODYDH)" in our study. Furthermore, [2] showed that machine learning models incorporating indicators of psychological and financial distress could be used for the early identification of students on academic probation, enabling timely and targeted interventions. This confirms the relevance of including socio-economic and emotional pressure indicators such as "Negative Impact of Home Environment (NegativeIH)" and "Pocket Money."

Despite this progress, significant gaps remain in the literature. Many studies have focused on general academic achievement rather than performance in specific, critical subjects like mathematics. A large portion of existing research has been conducted in developed countries or well-resourced environments, limiting its direct applicability to regions in sub-Saharan Africa. Finally, many advanced models lack the interpretability required for educators and policymakers to derive practical, actionable insights. This study aims to fill these gaps by applying and comparing both simple and sophisticated ML models to a comprehensive, context-specific dataset from Ghana. Our focus is not only on predictive accuracy but also on model interpretability, with the goal of providing clear, data-driven recommendations to improve mathematics education.

3 Methods

3.1 Data Source and Sample

Data were collected via a standardized questionnaire administered to 364 students from three Junior High Schools in the Ashanti Region of Ghana. The dataset comprises 26 variables covering student demographics (Gender, Age), socio-economic factors (Parent's Occupation, Family Financial Support, Pocket Money), study habits (Study Hours, Homework Frequency), and school environment (Class Size, Teacher Quality, Classroom Conduciveness). The primary outcomes of interest were the students' numerical mathematics score ('Mark', 0-100) and a binary grade category ('Grade': Pass/Fail, with a pass mark of 40).

3.2 Data Preprocessing

The raw data underwent several preprocessing steps to prepare it for machine learning modelling. Irrelevant identifier columns and columns with over 70% missing values were removed. Remaining missing values in categorical features were imputed using the mode.

Categorical variables were encoded using two methods:

- **Label Encoding:** Applied to ordinal variables where a natural order exists (e.g., 'Study Hours', 'Family Support').

- **One-Hot Encoding:** Applied to nominal variables to prevent the models from assuming an ordinal relationship (e.g., 'Gender', 'Parent's Occupation').

Finally, all numerical features were standardized using 'StandardScaler' from scikit-learn to have a mean of 0 and a standard deviation of 1, ensuring that algorithms sensitive to feature scale, such as SVM, could perform optimally.

3.3 Machine Learning Models

A suite of four supervised machine learning models was employed to address both regression and classification tasks.

- **Linear Regression:** A baseline model used to predict the continuous 'Mark' variable, assuming a linear relationship between predictors and the outcome.
- **Random Forest Regressor:** An ensemble model consisting of multiple decision trees, used to predict 'Mark'. It is robust to non-linear relationships and interactions between features.
- **Decision Tree Classifier:** A tree-based model used for the binary classification task of predicting 'Grade' (Pass/Fail). It is highly interpretable.
- **Support Vector Machine (SVM):** A powerful classification algorithm that finds an optimal hyperplane to separate classes. It was used to predict 'Grade'.

3.4 Model Training and Evaluation

The dataset was split into a training set (70%) and a testing set (30%). To optimize model performance and prevent overfitting, we performed hyperparameter tuning using 'GridSearchCV' with 5-fold cross-validation on the training data.

Model performance was assessed using standard metrics:

- **For Regression (Mark prediction):** Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2).
- **For Classification (Grade prediction):** Accuracy, Precision, Recall, and F1-Score. Confusion matrices were also analyzed to understand class-specific performance.

Feature importance was extracted from the trained Random Forest model to identify the most influential predictors of mathematics performance.

4 Results

4.1 Exploratory Data Analysis

Initial analysis of the data revealed key characteristics of the student sample. The distribution of marks was approximately normal, with most students scoring between 40 and 60. The dataset consisted of approximately 62.9% females and 37.1% males. The majority of students were in the 15-17 age group. Socio-economic indicators showed diversity, though most students reported receiving fair family support and belonged to middle-income households. Regarding study habits, 45.9% of students reported studying for 2 hours per day. Institutionally, most classes had 30-60 students, and the majority of students (52.2%) rated their teachers as excellent.

4.2 Model Performance: Regression

The Random Forest Regressor demonstrated substantially superior performance in predicting students' mathematics scores compared to the Linear Regression model. As shown in Table 1, the Random Forest explained 51% of the variance in marks ($R^2 = 0.51$), a significant improvement over the 16% explained by the linear model. The error metrics (MAE and RMSE) for the Random Forest were also considerably lower. Cross-validation confirmed the models' stability, with the Random Forest (CV RMSE = 14.09) again outperforming the Linear Regression (CV RMSE = 20.81), indicating better generalization to unseen data.

Table 1: Performance of Regression Models on Test Set

Model	MAE	RMSE	R^2	CV RMSE
Linear Regression	16.27	19.16	0.16	20.81
Random Forest Regressor	**10.71**	**14.68**	**0.51**	**14.09**

4.3 Model Performance: Classification

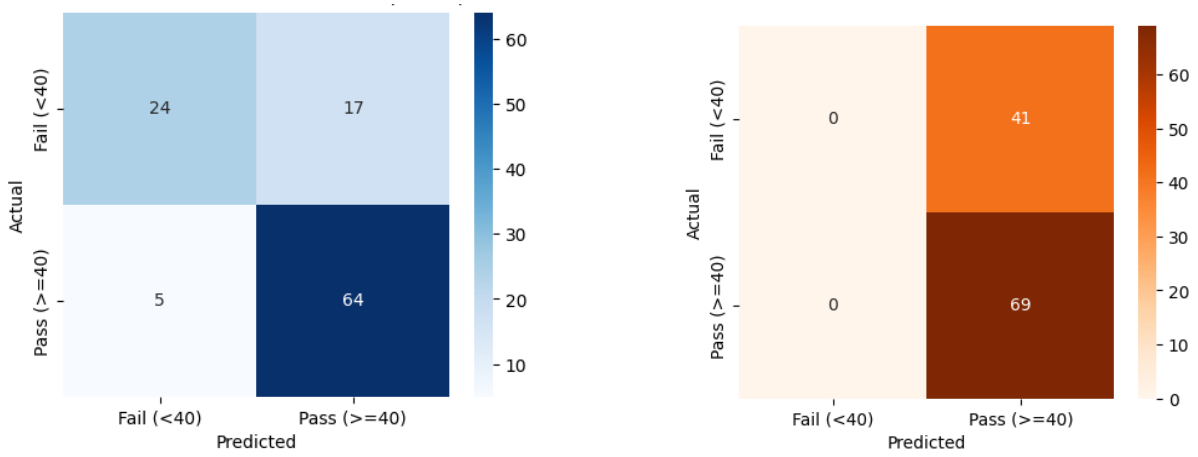
For the classification task of predicting whether a student would pass or fail, the Decision Tree classifier was the most effective model, achieving an overall accuracy of 80%. The SVM classifier performed poorly, with an accuracy of only 63%. As shown in Table 2, the SVM completely failed to identify any students in the "Fail" category (Precision and Recall of 0.00), likely due to the class imbalance in the dataset.

Table 2 provides a detailed breakdown of the models' performance. The Decision Tree, while highly effective at identifying passing students (Recall = 0.93), was less successful at identifying failing students (Recall = 0.59), correctly flagging only 59% of those at risk. The model's cross-validation accuracy was 76%, confirming its reliability.

Table 2: Classification Report for Decision Tree and SVM Models

Model	Class	Precision	Recall	F1-score	Support
Decision Tree	Fail (< 40)	0.83	0.59	0.69	41
	Pass (≥ 40)	0.79	0.93	0.85	69
	Accuracy			**0.80**	**110**
SVM	Fail (< 40)	0.00	0.00	0.00	41
	Pass (≥ 40)	0.63	1.00	0.77	69
	Accuracy			**0.63**	**110**

The analysis of the confusion matrices in Figure 1a and Figure 1b provides a more granular view of the classification performance. The Decision Tree correctly identifies 24 failing students but misclassifies 17 as passing. The SVM, however, misclassifies all 41 failing students.



(a) Confusion Matrix - DT (Tuned)

(b) Confusion Matrix - SVM (Tuned)

Figure 1: Confusion Matrices for Classification Models

4.4 Feature Importance

The feature importance analysis from the Random Forest model identified the top 10 factors that most significantly influence mathematics performance. Figure 2 illustrates that institutional and environmental factors are dominant. A detailed breakdown of the top predictors reveals a multi-faceted view of student performance.

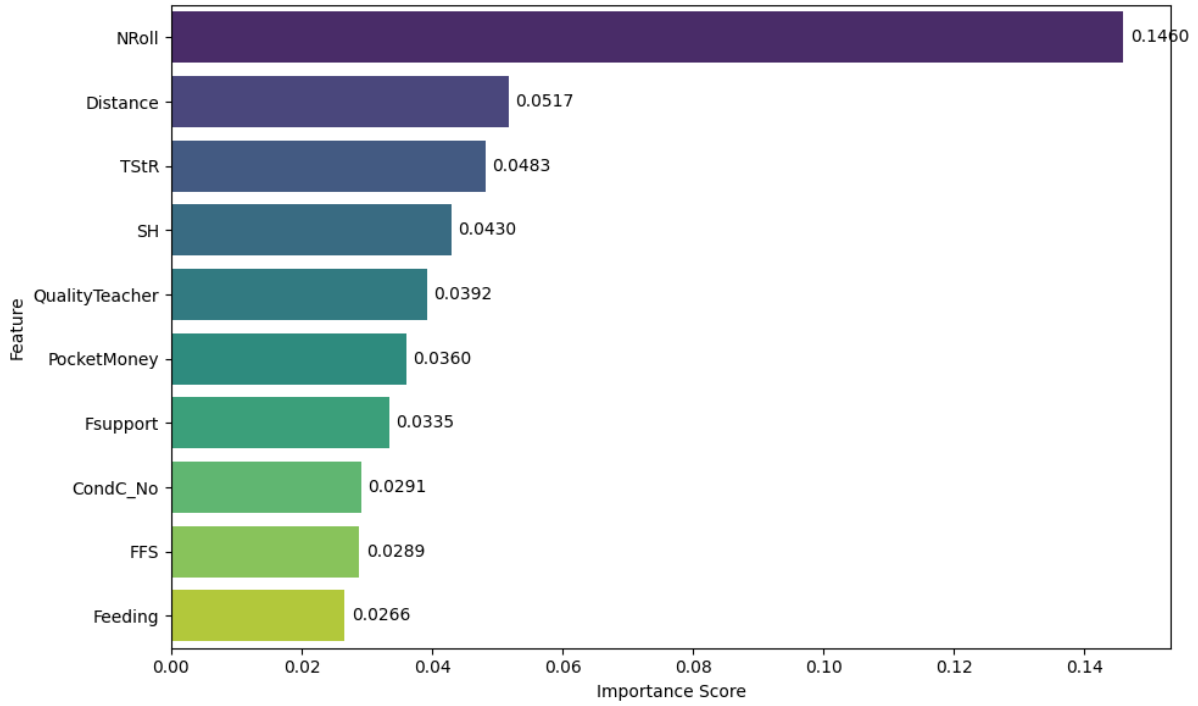


Figure 2: Top 10 Most Important Features

- **NRoll (Class Size):** The most critical predictor. Larger classes may dilute individual teacher attention and strain resources.
- **Distance:** The second most important factor. Long travel times can lead to fatigue and reduce available study time.

- **TStR (Teacher's Rating):** The teacher's academic assessment of the student. This suggests that teachers' perceptions and feedback are strongly correlated with outcomes.
- **SH (Study Hours):** A direct measure of student effort, highlighting the importance of time-on-task outside the classroom.
- **QualityTeacher:** The student's perception of teaching effectiveness is a key driver of their performance.
- **PocketMoney & FFS:** These serve as proxies for socio-economic status, indicating that financial stability impacts a student's ability to focus on academics.
- **Fsupport (Family Support):** The level of academic encouragement from family is a significant factor.
- **CondC_No (Non-Conducive Classroom):** A negative indicator for classroom environment, showing that factors like noise or poor ventilation hinder learning.
- **Feeding:** Proper nutrition is crucial for cognitive function and concentration.

5 Discussion

This study successfully applied supervised machine learning to model the mathematics performance of Ghanaian Junior High School students, yielding several key insights. The superior performance of the Random Forest model over Linear Regression underscores the non-linear and interactive nature of the factors influencing academic achievement, consistent with findings by [1]. Simple linear models are insufficient to capture the complex interplay between student, family, and school variables. This aligns with critiques by [23], who noted that linear models often fail to capture educational complexities.

The feature importance analysis provides compelling, data-driven evidence for educational policy. The prominence of **class size (NRoll)** as the top predictor aligns with extensive educational research suggesting that smaller classes can facilitate more individualized attention and better learning outcomes [7]. Similarly, logistical factors like **distance to school** impose a tangible burden on students, potentially leading to fatigue and reduced study time. This aligns with the broad understanding that environmental and logistical factors can influence educational outcomes.

The strong influence of teacher-related variables (**Teacher's Rating**, **QualityTeacher**) reinforces the critical role of educators in student success [1, 24]. Student perceptions of teacher quality and the feedback they receive are directly linked to their performance. This highlights the need for continuous professional development and support for teachers.

Socio-economic factors, represented by **PocketMoney**, **Family Support**, and **Family Financial Status**, also emerged as significant predictors. This confirms that a student's home environment and economic stability are inextricably linked to their academic life, a finding consistent with previous studies [2, 7].

The classification results, particularly the Decision Tree's ability to identify at-risk students with reasonable accuracy, suggest the practical utility of these models. Schools could potentially use such data-driven tools for early identification and targeted intervention, providing support to students before they fall significantly behind. However, the SVM's failure highlights the challenge of class imbalance in educational datasets and the importance of selecting appropriate algorithms or using techniques like SMOTE to mitigate this issue.

6 Conclusion

This study applied supervised machine learning techniques to model and predict mathematics performance among Junior High School students in Ghana. Using data from 364 students across demographic, socio-economic, behavioural, and school-related variables, the analysis identified critical factors influencing performance.

The Random Forest Regressor outperformed Linear Regression in predicting continuous scores ($R^2 = 0.51$ vs. 0.16), while the Decision Tree classifier surpassed the SVM in predicting pass/fail outcomes (80% vs. 63% accuracy). Feature importance analysis revealed that class size, distance to school, teacher's rating, study hours, and teacher quality were the most influential predictors. These findings confirm that both institutional and household factors jointly shape learning outcomes.

Limitations The findings of this study are based on a localized dataset of 364 students and self-reported data, which may limit generalisability and introduce potential bias. The predictive power of the best model ($R^2 = 0.51$) may indicate that a significant portion of performance variance remains unexplained, suggesting that other unmeasured variables also play a role.

Future Work Future research should expand the scope by incorporating larger and more diverse datasets, ideally collected across multiple regions or longitudinally, to enhance generalisability. From a methodological perspective, advanced ensemble techniques such as Gradient Boosting and XGBoost, as well as class imbalance handling methods like SMOTE, could further improve predictive accuracy.

At the policy and practice level, reducing class sizes, improving teacher quality, and addressing logistical barriers like long travel distances are important areas for intervention. Support measures such as school feeding and financial aid may also help reduce performance gaps tied to poverty. Including feedback from teachers and parents would enhance the quantitative data, offering a better understanding of the factors affecting mathematics achievement.

This study shows the potential of machine learning in delivering data-driven insights for educational improvement. It highlights practical steps for both research and practice in mathematics education in Ghana.

References

[1] Aksu, N., Aksu, G., & Saracaloglu, A. S. (2022). Prediction of the factors affecting PISA mathematics literacy of students from different countries by using data mining methods. *International Electronic Journal of Elementary Education, 14*(5), 613–629.

[2] Al-Alawi, L., Shaqsi, J. A., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: The case of college students on academic probation. *Education and Information Technologies*, 1-24.

[3] Arpa, T., & Çavur, M. (2024). A comparative analysis of machine learning techniques to explore factors affecting mathematics success in developing countries: Turkey, Mexico, Thailand, and Bulgaria case studies. *Journal of Information Systems and Management Research, 6*(2), 24–36.

[4] Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32.

[5] Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

[6] Brownlee, J. (2020). *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

[7] Chen, S., & Ding, Y. (2023). A machine learning approach to predicting academic performance in Pennsylvania's schools. *Social Sciences, 12*(3), 118.

[8] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*, 273–297.

[9] Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

[10] El Guabassi, I., Bousalem, Z., Marah, R., & Qazdar, A. (2021). Comparative analysis of supervised machine learning algorithms to build a predictive model for evaluating students' performance. *International Journal of Online and Biomedical Engineering, 17*(2), 90–105.

[11] Elrahman, A. A., Soliman, T. H., Taloba, A. I., & Farghally, M. F. (2022). A predictive model for student performance in classrooms using student interactions with an etextbook. *arXiv preprint arXiv:2203.03713*.

[12] Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. O'Reilly Media, Inc.

[13] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*(Mar), 1157–1182.

[14] 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, p. 032019). IOP Publishing.

[15] Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). *The elements of statistical learning*. Springer.

[16] Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning, 14*(14), 92–104.

[17] Jin, X. (2023). Predicting academic success: machine learning analysis of student, parental, and school efforts. *Asia Pacific Education Review*, 1-13.

[18] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science, 1*(2), 111–117.

[19] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-hill.

[20] Lee, J. E., Jindal, A., Patki, S. N., Gurung, A., Norum, R., & Ottmar, E. (2023). A comparison of machine learning algorithms for predicting student performance in an online mathematics game. *Interactive Learning Environments, 32*(9), 5302–5316.

[21] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news, 2*(3), 18–22.

[22] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR), 9*(1), 381–386.

[23] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[24] Mosia, M., Egara, F. O., Nannim, F. A., & Basitere, M. (2024). Factors influencing students' performance in university mathematics courses: A structural equation modelling approach. *Education Sciences, 15*(2), 188.

[25] Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237.

[26] Oppong, S. O. (2023). Predicting students' performance using machine learning algorithms: A review. *Asian Journal of Research in Computer Science, 16*(3), 128–148.

[27] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128–138.

[28] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning, 1*, 81–106.

[29] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160.

[30] Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. *International Journal of Modern Education and Computer Science, 13*(1), 1–10.

[31] Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

[32] Shetty, S. H., Shetty, S., Singh, C., & Rao, A. (2022). Supervised machine learning: Algorithms and applications. In *Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools and Applications* (pp. 1–16). Springer.

[33] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

[34] Wang, Y., Ding, A., Guan, K., Wu, S., & Du, Y. (2021). Graph-based ensemble machine learning for student performance prediction. *arXiv preprint arXiv:2112.07893*.

[35] Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*.

[36] Yauri, R. A., Suru, H. U., Afrifa, J., & Moses, H. G. (2022). A machine learning approach in predicting student's academic performance using artificial neural network. *Journal of Computational and Cognitive Engineering, 3*(2), 203–212.