University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR)

ISSN: 2714-3627

A Journal of the Department of Computer Science, University of Ibadan, Ibadan, Nigeria

Volume 14 No. 1, June, 2025

journals.ui.edu.ng/uijslictr
http://uijslictr.org.ng/

University of Ibadan
Journal of Science and Logics in ICT Research

Predicting Student Academic Performance Using a Scalable Regression-Based Data Mining Approach

[1]Adejumo A., [2]Woods, N. C., [3]✉ Ojo, A. K.

[1,2,3]University of Ibadan, Department of computer Science, Ibadan, Nigeria
[1]bolamf@gmail.com; [2]chyn.woods@gmail.com; [3]adebolak.ojo@gmail.com

Abstract
Predicting student academic performance is a key tool for supporting academic planning and identifying those who may need extra help. This study develops a regression-based model aimed at forecasting academic outcomes among students at the University of Ibadan, Nigeria. Data were collected from 92 departments over a three-year period, covering both academic records and non-academic factors. After data preparation—which involved cleaning, feature selection, and encoding—three regression techniques were applied: Stochastic Gradient Descent (SGD), Gradient Boosting Machine (GBM), and Extra Trees Regressor (ETR). Among these, the ETR model gave the most accurate predictions, based on performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). The use of loss functions such as Huber further improved the model's ability to handle outliers. The findings show that this model can help pinpoint students at risk of poor academic performance and support better decisions in academic advising, resource planning, and policy implementation.

Keywords: Student Performance Prediction, Machine Learning in Education, Educational Data Mining, Predictive Analytics, Regression Analysis

1. Introduction

Improving student academic outcomes and reducing dropout rates remain central challenges in higher education today. Institutions increasingly turn to data-driven methods to identify students who may be at risk of underperforming. Early detection allows for timely academic interventions that support student retention and success. This study proposes a regression-based model to forecast academic performance using variables such as socio-economic background, attendance, and previous academic results. The approach aims to strike a balance between accuracy and interpretability, making it suitable for practical use in academic institutions.

Focusing on students at the University of Ibadan, Nigeria, the research addresses the growing need for evidence-based decision-making in African universities. Recent work by [1] emphasized the value of learning analytics for identifying gaps in academic progress and improving institutional responsiveness. By ensuring the model is both scalable and adaptable, this study contributes to ongoing efforts to integrate predictive analytics into university systems for more informed student support.

2. Related Works

Several studies have explored methods for predicting student performance using machine learning. Earlier research relied on classification algorithms such as decision trees, support vector machines, and k-nearest neighbours [2, 3]. While these models provided some success, they often struggled with larger datasets and lacked interpretability—an important requirement for use in academic planning.

More recent research has shown that including non-academic variables such as student background, class attendance, and behaviour improves model accuracy. For example, Arulmozhi *et. al.* [4] and Elango *et. al.* [5] found that adding these features helped improve performance prediction. Nuñez *et. al.* [6] also stressed the importance of using a broader range of data points to capture key differences in student learning.

Adejumo A., Woods, N. C., Ojo, A. K. (2025). Predicting Student Academic Performance Using a Scalable Regression-Based Data Mining Approach. *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 14 No. 1, pp. 98 – 104.

Despite these advances, some problems remain. Many studies are limited by small sample sizes or imbalanced data. Others use complex models without offering clear explanations for their predictions [7, 8]. Romero and Ventura [9] argued that predictive models in education need to be both accurate and easy to interpret, especially when used to guide real academic decisions. More recently, Van der *et. al.* [10] emphasized transparency as essential when models are used to influence student support strategies.

Some work has also been done in Nigerian institutions. Ojo and George [11] applied clustering methods to study student admission data at the University of Ibadan. Others, such as Alabi *et. al.* [12], Oyefolahan *et. al.* [13], and Girma [14], highlighted the need for better predictive systems to address poor performance and student dropouts.

To overcome the weaknesses of single-model approaches, researchers have increasingly turned to ensemble methods. Studies by Roslan and Chen, [15], Khan and Ghosh, [16], and Costa *et. al.* [17] have shown that combining models can produce more stable and accurate results. These findings support the use of ensemble regression methods in this study.

This research contributes to existing work by applying and comparing three regression models—Stochastic Gradient Descent (SGD), Gradient Boosting Machine (GBM), and Extra Trees Regressor (ETR)—on student data from a Nigerian university. The aim is to identify which approach provides the best balance of accuracy, reliability, and ease of interpretation for practical academic use.

3. Methodology

This study proposes a scalable regression-based model to predict student academic performance, combining advanced data mining techniques with domain-driven feature selection. As illustrated in Figure 1, the methodological workflow spans data acquisition, preprocessing, model development, and evaluation.

3.1 Data Acquisition and Scope

The dataset was obtained from the University of Ibadan, covering three academic sessions from 2020/2021 to 2022/2023. It consists of 20,870 student records, representing 92 departments across 17 faculties. For consistency and representativeness, three core undergraduate courses were selected: GES 101 (General Studies), CHE 156 (Chemistry), and ZOO 114 (Zoology). Extracted features include Department, Course Level, Semester, Academic Session, Continuous Assessment (CA) scores, Examination scores, and Final Course Results—carefully curated for predictive modelling.

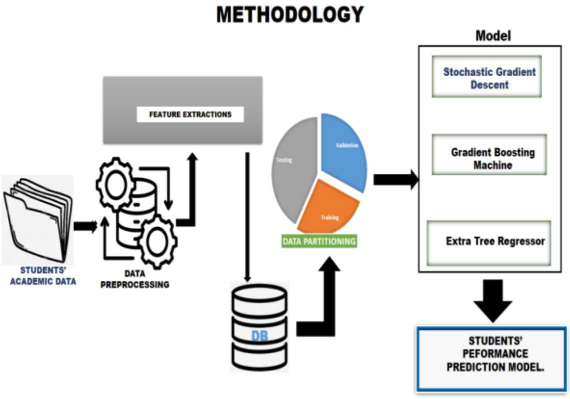

Figure 1: Methodology workflow

3.2 Data Preprocessing

Data preprocessing entailed cleaning inconsistencies, encoding categorical variables numerically, and engineering features that enhance the model's predictive capacity. Categorical fields such as department and course level were transformed using label encoding. Numerical features were standardized to ensure consistent scale. The complete dataset was partitioned into training (80%), validation (10%), and test (10%) sets to ensure reliable evaluation and minimize overfitting.

3.3 Predictive Modeling Algorithms

Three regression models were employed to capture different patterns in the data: Stochastic Gradient Descent (SGD), Gradient Boosting Machine (GBM), and Extra Trees Regressor (ETR). Each algorithm offers distinct advantages in terms of scalability, interpretability, and predictive strength.

3.3.1 Stochastic Gradient Descent (SGD)

SGD is a linear optimization algorithm that updates model parameters iteratively using one training example at a time. This method is computationally efficient for large datasets and suitable for real-time model updates. The steps involved include:

1. Prediction of the output:

$$\hat{y}_i = w.x_i + b \qquad (1)$$

2. Calculation of the loss of the i_{th} sample:

$$L_i = \frac{1}{2}(y_i - \hat{y}_i)^2 \qquad (2)$$

3. Computation of the gradients of the loss with respect to the weights and bias:

$$\frac{\partial L_i}{\partial w} = -(y_i - \hat{y}_i)x_i \qquad (3)$$

$$\frac{\partial L_i}{\partial b} = -(y_i - \hat{y}_i)x_i \qquad (4)$$

4. Parameter updates using the gradients and learning rate η:

$$w \leftarrow w - \eta \frac{\partial L_i}{\partial w} \qquad (5)$$

$$b \leftarrow b - \eta \frac{\partial L_i}{\partial b} \qquad (6)$$

Where η is the learning rate controlling update magnitude. This iterative process continues until convergence.

3.3.2 Gradient Boosting Machine (GBM)

GBM is an ensemble technique that builds decision trees sequentially, with each tree aiming to correct the residuals of the previous model. This results in a strong predictive model through a series of small improvements. The process begins with an initial prediction (e.g., mean target value). Residuals are calculated, and a weak learner is fit to these residuals. This learner's predictions are then added to the existing model with a learning rate v to control influence:

$$F_m(x) = F_0(x) + v \sum_{m=1}^{M} hm(x) \qquad (7)$$

Training continues iteratively until a specified number of boosting rounds is reached or performance improvement becomes marginal. GBM is especially effective for reducing both bias and variance in predictive tasks.

3.3.3 Extra Trees Regressor (ETR)

ETR is a non-parametric ensemble method that constructs multiple unpruned decision trees, introducing randomness in the choice of split thresholds and features. This additional randomness enhances generalization while speeding up training.

Key steps include:

Dataset Representation:

$$D = \{(x_1, y_1), \ldots, (x_n, y_n)\} \qquad (8)$$

where each feature vector $x_i \in \mathbb{R}^D$ and the corresponding target $y_i \in \mathbb{R}$.

Bootstrap Sampling:

Random subsets of training data are sampled with replacement for each tree.

Random Splits:

Instead of finding the optimal split, ETR selects split points randomly from feature subsets, reducing training time.

Prediction:

Final output is the average of all tree predictions:

$$\hat{y} = \frac{1}{t}\sum_{t=1}^{T} f_t(x) \qquad (9)$$

Validation: Final predictions (ŷ) are obtained by averaging the outputs of all trees in the ensemble. Performance is evaluated on a validation set, using metrics such as MSE.

$$D_{val} = \{(x_{1val}, y_{1val}), (x_{2val}, y_{2val}), \ldots, (x_{mval}, y_{mval})\} \qquad (10)$$

ETR's strength lies in its robustness and low variance, making it suitable for complex datasets with noise.

3.4 Model Evaluation and Implementation Tools

Models were evaluated using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2). These metrics provide insights into prediction accuracy and model fit.

Hyperparameters such as tree depth, number of estimators, and learning rates were optimized using grid search combined with K-fold cross-validation. The models were developed using Python and popular data science libraries:

Scikit-learn: Modelling and evaluation

NumPy & Pandas: Data manipulation and processing

Matplotlib & Seaborn: Visualization

4. Results and Discussion

4.1 Model Performance and Feature Importance

An analysis of feature importance, as shown in Figure 2, reveals the key factors that influence student academic outcomes. Examination scores had the strongest impact, indicating their critical role in final grade determination. The CA followed as another major contributor, underscoring the importance of consistent academic performance during the semester.

In contrast, variables such as Semester and Course Level had relatively low importance scores, suggesting that their influence on academic results is minimal.

Insights from Feature Engineering

The model was trained on data from academic sessions spanning 2020 to 2023. Although the importance of core predictors like Exam and CA remained relatively stable, slight variations were observed across sessions, especially in contextual features. Variables like Semester, which consistently contributed less to the model's accuracy, could be removed or merged to reduce dimensionality without losing valuable predictive information.

Recommendations:

Maintain a strong focus on Exam and CA scores in performance prediction models.

Improve prediction quality by integrating new variables such as attendance records, study behaviour, or participation in academic support programs.

Remove features with low predictive power to simplify the model and enhance its interpretability without compromising accuracy.

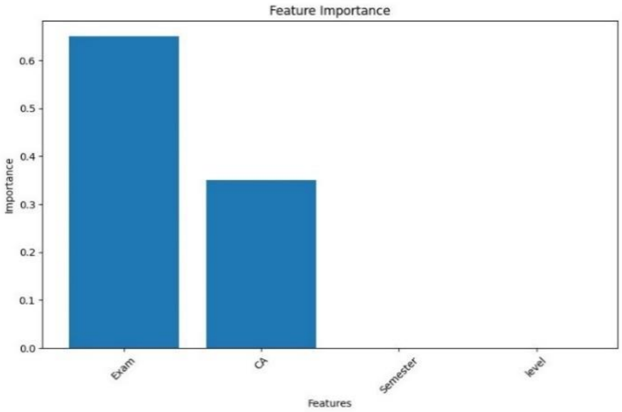

Figure 2: Selected and engineered key features

4.2 Distribution of Key Features

Level Distribution:

Figure 3 shows that most students in the dataset were at the 100 level. The number declines at higher levels (200 to 600), likely due to factors such as direct entry admissions, repeated courses, or course carryovers. This skewed distribution could affect academic planning and student progression analysis.

Continuous Assessment (CA):

The histogram of CA scores approximates a normal distribution, where most scores fall within an average range. This makes CA a reliable metric for gauging ongoing academic performance. Students at the extremes of this distribution may require additional academic support.

Examination Scores:

The distribution of exam scores is also close to normal, with scores concentrated around the mean. This supports the finding that exams are a dependable indicator of student achievement across the board.

Final Course Results:

Result scores (ranging from 0 to 100) show a normal bell-shaped curve, peaking around 60, with approximately 1,200 students in that range. This consistency across distributions adds credibility to the model's predictions.

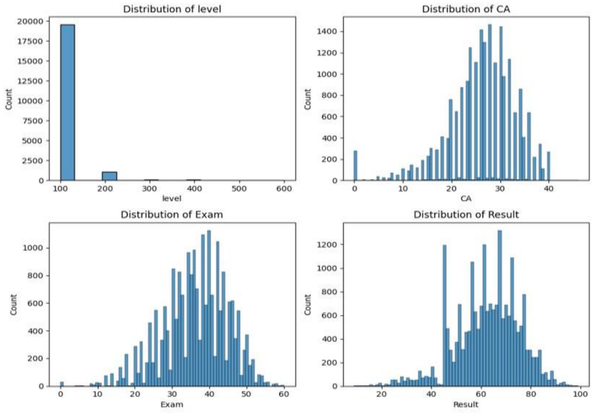

Figure 3: Histograms showing the distribution of various variables used to predict student performance

4.3 Top-Performing Departments

Figure 4 compares departmental averages for student performance across 92 departments. The highest-performing departments—such as Medicine, Pharmacy, and Communication—consistently recorded above-average results. This suggests that departmental context can influence academic performance and should be considered when developing predictive models.

4.4 Effect of Epochs on Model Performance

The number of training epochs can significantly affect model accuracy. An epoch refers to one full cycle through the training dataset. In this study, three machine learning models were evaluated: Stochastic Gradient Descent (SGD), Gradient Boosting Machine (GBM), and Extra Trees Regressor (ETR).

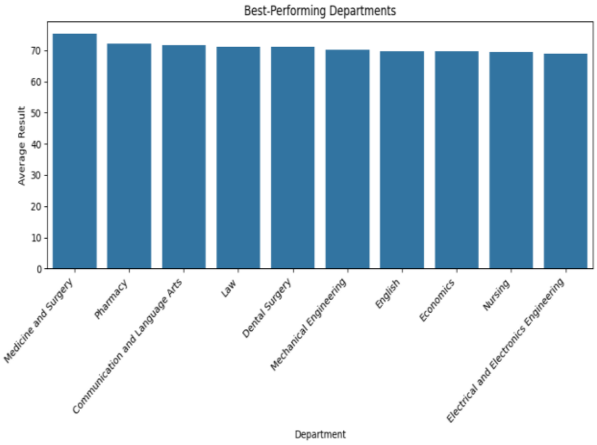

Figure 4: Prediction of the best high-ranking departments

a) Stochastic Gradient Descent (SGD)

Figure 5 shows that the model’s R² score increased rapidly at the start of training and plateaued at about 0.996 after several epochs. This indicates convergence, beyond which further training yielded little improvement. Despite its stochastic behaviour, the model achieved a high level of accuracy.

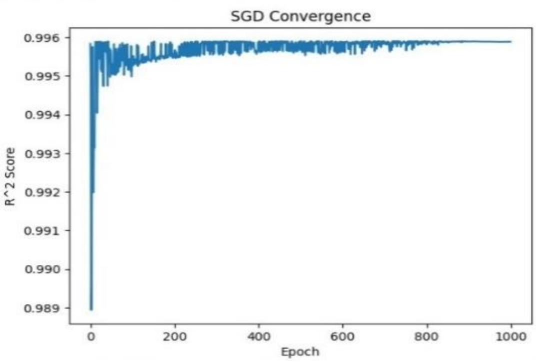

Figure 5: Epoch predicting the performance of SGD

b) Gradient Boosting Machine (GBM)

Figure 6 demonstrates that GBM’s Mean Squared Error (MSE) decreased from 0.425 to around 0.385 before levelling off near epoch 100. This pattern reflects good learning behaviour, though additional training brought diminishing returns.

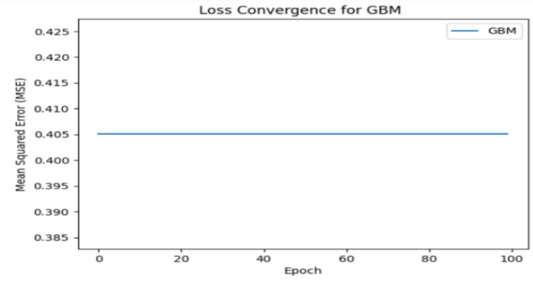

Figure 6: Epoch used to predict the performance of GBM

c) Extra Trees Regressor (ETR)

As shown in Figure 7, ETR achieved rapid convergence, with MSE values consistently between 0.124 and 0.132. This stability suggests that ETR can achieve accurate predictions with relatively few epochs and is less sensitive to overfitting.

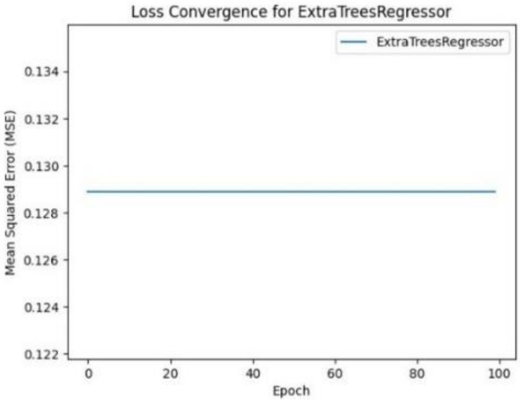

Figure 7: Epoch predicting the performance of ETR

4.5 Model Evaluation and Comparative Performance

The predictive performances of the three models were evaluated using MSE, RMSE, and R² (Table 1). The results are summarized as follows:

Table 1: Three machine learning models' efficacy or efficiency

Model	MSE	RMSE	R²
SGD	0.1441	0.3797	0.9991
GBM	0.2935	0.5418	0.9982
ETR	0.0655	0.2559	0.9996

Among the three, ETR delivered the best overall performance, achieving the lowest error rates and highest predictive accuracy. SGD followed closely, showing robust performance and resistance to outliers. GBM, while reliable, displayed higher error margins, possibly due to overfitting or tuning limitations.

Figure 8 highlights high error variability using squared loss in SGD—indicating the need for loss function tuning. Figure 9 (SGD with Huber Loss) shows tight clustering around the diagonal, enhancing robustness against outliers. Figure 10 (GBM) also displays reliable prediction accuracy.

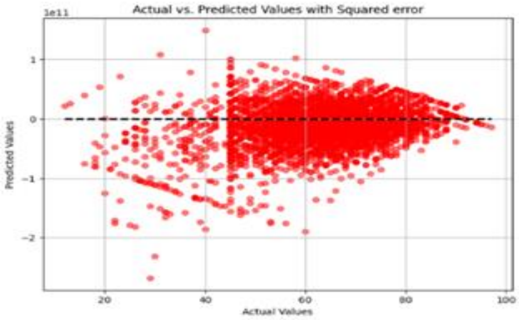

Figure 8: Prediction of SGD using Squared loss function

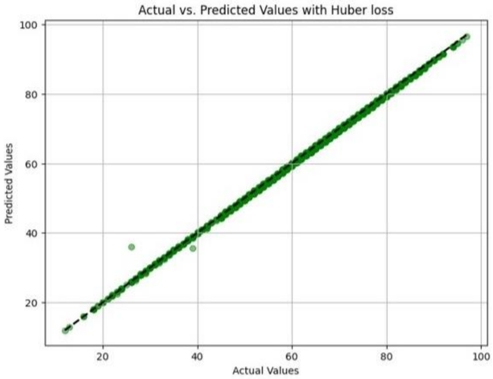

Figure 9: SGD Prediction with Huber loss function

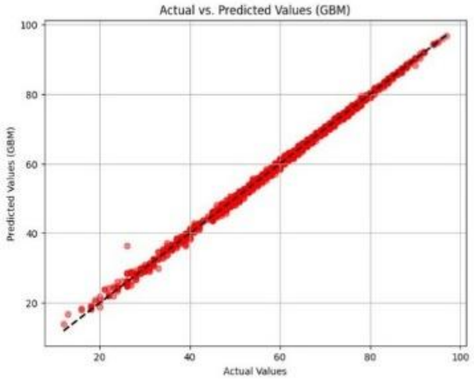

Figure 10: GBM Prediction

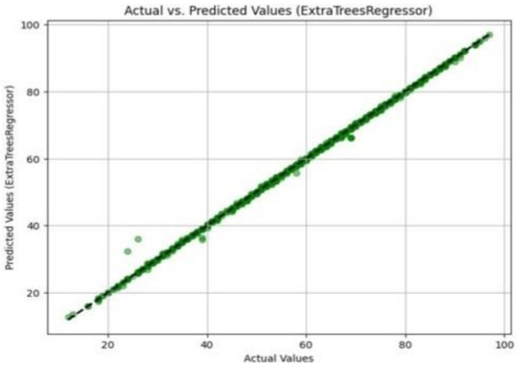

Figure 11: ETR prediction

Figure 11 (ETR): Exhibits the closest alignment with the ideal 45° line, confirming its precision. Figure 12: Reinforces the advantage of using custom loss functions for optimization.

5. Conclusion

This study explored the use of regression models to forecast student academic performance, applying three techniques: Extra Trees Regressor (ETR), Stochastic Gradient Descent (SGD), and Gradient Boosting Machine (GBM). Among the three, ETR produced the most accurate results, reflected in its low Mean Squared Error (0.0655), low Root Mean Squared Error (0.2559), and a high R^2 score (0.9996). These results suggest that ETR is particularly effective in tracking academic patterns with minimal error.

Although SGD showed resilience to outliers and GBM performed well in capturing non-linear trends, both models required more fine-tuning to match ETR's performance. ETR not only delivered reliable predictions but also did so efficiently, making it a practical option for institutions managing large datasets.

The outcomes of this research point to the value of choosing models that strike a balance between accuracy, ease of interpretation, and scalability. The ETR model, in particular, offers strong potential for early identification of students who may be at risk, enabling educators and administrators to intervene in time. Future studies could enhance the model by incorporating behavioural and contextual data to improve its application across different learning environments.

References

[1] Ogundaini, O., & Mlitwa, N., "Using learning analytics to improve teaching and learning in Sub-Saharan African universities: A policy and practice perspective.," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 1–18, 2023.

[2] Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshmawi, A. A., ... & Ashraf, I., "Role of convolutional features and machine learning for predicting student academic performance from MOODLE data," *PLOS ONE*, vol. 18, no. 11, p. e0293061, 2023.

[3] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, 2022.

[4] Arulmozhi, P., Hemavathi, N., Rayappan, J. B. B., & Raj, P., "ALRC: A novel adaptive linear regression based classification for grade based student learning using radio frequency

identification," *Wireless Personal Communications,* vol. 112, no. 4, p. 2091–2107, 2020.

[5] Elango, S., Natarajan, E., Varadaraju, K., Abraham Gnanamuthu, E. M., Durairaj, R., Mohanraj, K., & Osman, M. A., "Extreme gradient boosting regressor solution for defy in drilling of materials," *Advances in Materials Science and Engineering,* p. 1–13, 2022.

[6] Nuñez, H., Maldonado, G., & Astudillo, C. A., "Semi-supervised regression based on tree SOMs for predicting students' performance," in *2018 International Joint Conference on Neural Networks (IJCNN),* Rio de Janeiro, Brazil, 2018.

[7] Alyahyan, E., & Düştegör, D., "Predicting academic success in higher education: Literature review and best practices," *International Journal of Educational Technology in Higher Education,* vol. 17, no. 1, 2020.

[8] Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., ... & Liao, S. N., "Predicting academic performance: A systematic literature review," Larnaca, Cyprus, 2018.

[9] Romero, C., & Ventura, S., "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews,* vol. 40, no. 6, p. 601–618, 2010.

[10] Van der Wal, O., Bachmann, D., Leidinger, A., van Maanen, L., Zuidema, W., & Schulz, K., "Undesirable biases in NLP: Addressing challenges of measurement," *Journal of Artificial Intelligence Research (JAIR),* vol. 79, p. 1–40, 2024.

[11] Ojo, A. K. and George, A. E, "Optimal Clustering Algorithm for Knowledge Discovery in University of Ibadan Post Unified Tertiary Matriculation Examination," in *Proceedings of 4th Biennial Conference, Transition from Observation to Knowledge to Intelligence.,* Ibadan, Nigeria, 2021.

[12] Alabi, I. O., Alabi, S. O., & Adeyemo, A. B., "Academic performance prediction for success rate improvement in higher institutions of learning: An application of data mining classification algorithms," *Journal of Educational Data Mining (JEDM),* vol. 10, no. 2, p. 45–60, 2018.

[13] Oyefolahan, I. O., Idris, S., Etuk, S. O., & Alabi, I. O., "Academic performance prediction for success rate improvement in higher institutions of learning: An application of data mining classification algorithms," *https://jedm.educationaldatamining.org/index.php/JEDM/article/view/220,* vol. 10, no. 2, p. 45–60, 2018.

[14] S. Girma, "Developing a Predictive Model to Determine Higher Education Students' Academic Status Using Data Mining Technology," Addis Ababa, Ethiopia, 2019.

[15] Roslan, M. B., & Chen, C., "Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015–2021)," *International Journal of Emerging Technologies in Learning (iJET),* vol. 17, no. 5, p. 147–179, 2022.

[16] Khan, A., & Ghosh, S. K., "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Education and Information Technologies,* vol. 26, no. 1, p. 205–240, 2021.

[17] Costa, L. A., Pereira Sanches, L. M., Rocha Amorim, R. J., Nascimento Salvador, L. D., & Santos Souza, M. V. D., "Monitoring academic performance based on learning analytics and ontology: A systematic review," *Informatics in Education,* vol. 19, no. 3, p. 361–397, 2020.

[18] Yin, C., Tang, D., Zhang, F., Tang, Q., Feng, Y., & He, Z., "Students learning performance prediction based on feature extraction algorithm and attention-based bidirectional gated recurrent unit network," *PLOS ONE,* vol. 18, no. 10, p. e0286156, 2023.

[19] Grebla, H. A., Rusu, C. V., Sterca, A., Bufnea, D., & Niculescu, V., "Recommendation System for Student Academic Progress," in *14th International Conference on Agents and Artificial Intelligence (ICAART 2022),* 2022.

[20] Ghosh, S. K., Ghosh, A., & Das, S., "Student performance analysis and prediction in classroom learning: A review of educational data mining studies," *Computers & Education,* vol. 150, 2020.