ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2026; 5; 1-34

Fairness-Aware Machine Learning for Ethical Student Performance Prediction in Education

[a]Musa Adekunle Ayanwale , **[b]Ijeoma Joy Chikezie***

[a]University of Pretoria, Groenkloof, 0002, Pretoria, South Africa
[b]National Institute for Nigerian Languages, Aba, Abia State, Nigeria

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into education has transformed how student performance is predicted and monitored. Despite these advances, concerns regarding fairness, transparency, interpretability, and potential demographic bias remain significant challenges in educational prediction systems. This study developed ethically aligned and interpretable ML models for predicting student academic performance using only behavioural engagement and academic context variables. The open-access xAPI-Edu-Data dataset containing 480 student records was obtained from Kaggle. Four supervised algorithms, Multinomial Logistic Regression, Decision Tree, Random Forest, and XGBoost, were implemented in Python 3.11 using Scikit-learn, SHAP, and LIME frameworks. To minimise data leakage, all preprocessing operations, including standardisation and categorical encoding, were embedded within a Scikit-learn Pipeline fitted exclusively on the training folds. A stratified train/validation/test split (70%/15%/15%) with random_state = 42 was employed, while hyperparameters were optimised using five-fold cross-validation on the training and validation sets only. Model performance was evaluated using accuracy, macro F1-score, ROC-AUC, Brier Score, and Expected Calibration Error (ECE). Results showed that Random Forest achieved the best overall performance and calibration, while Logistic Regression provided superior interpretability. Across all models, visited_resources and raisedhands consistently emerged as the strongest predictors of academic achievement, emphasising the importance of student engagement behaviours. The study demonstrates that fairness-by-design prediction systems that exclude demographic variables can support equitable and actionable early-warning interventions. The novelty of the study lies in its ethically grounded and deployment-oriented framework for interpretable educational prediction in resource-constrained contexts.

Keywords: explainable artificial intelligence, student performance prediction, fairness-aware machine learning, learning analytics, educational data mining

* Corresponding author.

National Institute for Nigerian Languages, Aba, Abia State, Nigeria, e-mail: drijeomajchikezie@gmail.com

To cite this article:

Ayanwale, M. A., & Chikezie, I. J. (2026). Fairness-aware machine learning for ethical student performance prediction in education. *Journal of Computer Adaptive Testing in Africa, 5*, 1-34. https://doi.org/10.71291/68teb107

Received: 10 April 2026
Accepted: 25 June 2026
Published: 21 July 2026

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) are increasingly changing the way educational systems operate, influencing not only how students learn but also how institutions assess performance, monitor engagement, and make academic decisions. Across schools and universities, AI-supported systems are now being used to analyse learner data, identify students who may be academically at risk, automate aspects of assessment, and support evidence-based institutional planning (Ahmed et al., 2024; Wang et al., 2024). These developments reflect a broader movement toward data-informed education, where predictive insights can help educators intervene earlier, personalise instructional support, and improve overall learning outcomes (Vieriu & Petrea, 2025). In many respects, educational systems are shifting from reactive approaches, where intervention occurs only after poor performance becomes visible, toward proactive models capable of identifying potential learning difficulties before they escalate.

Despite these promising developments, the growing use of AI in education has also generated serious ethical and pedagogical concerns. One of the most persistent challenges relates to fairness. Predictive algorithms trained on historical educational data may unintentionally reproduce inequalities already embedded within those datasets, particularly when demographic variables such as gender, nationality, or socio-economic status are included in model training (Bulut et al., 2024; Wang et al., 2024). Consequently, systems designed to support students may instead reinforce existing disadvantages. Another major concern involves transparency. Many advanced ML models, especially ensemble and deep learning systems, often function as "black boxes," producing highly accurate predictions without clearly explaining how those predictions are generated (Linardatos et al., 2021; Gunasekara & Saarela, 2025). While such systems may offer strong predictive performance, their opacity creates difficulties for educators who must justify algorithm-informed decisions in real educational settings. As a result, there is increasing recognition that educational AI systems should not be evaluated solely based on predictive accuracy. They must also be interpretable, ethically defensible, transparent, and aligned with educational values.

Within the area of student performance prediction, research has expanded considerably over the last decade. Educational data mining and learning analytics studies have applied ML algorithms to predict academic achievement, course completion, learner engagement, and dropout risk using behavioural logs, assessment records, demographic profiles, and institutional datasets (Arizmendi et

ISSN: 2790-4407 www. jocatia.acata.org

al., 2023). Ensemble approaches such as Random Forest and boosting algorithms frequently outperform traditional statistical models because they can capture complex and non-linear relationships among educational variables more effectively (Ayulani et al., 2023). However, the increasing sophistication of these predictive systems has often come at the expense of interpretability. Even highly accurate models may struggle to explain why specific predictions are made, particularly when differentiating between students with moderate and high achievement levels (Deep Interpretable Ensembles, 2022). This creates a central challenge within educational AI research: how to develop systems that achieve strong predictive performance while remaining transparent, trustworthy, and educationally meaningful.

This study responds to this challenge by proposing an ethically grounded and explainable framework for student performance prediction. Unlike many previous studies that depend heavily on demographic variables to improve predictive accuracy, the present study focuses primarily on behavioural engagement and academic context variables. Specifically, the study examines indicators such as visited_resources, raisedhands, announcements_view, discussion participation, topic, stage level, grade category, and semester. These variables were selected because they are both pedagogically actionable and institutionally meaningful. Educators can intervene when students demonstrate reduced engagement with learning materials or classroom participation, whereas demographic characteristics are largely immutable and ethically sensitive. By prioritising behavioural engagement indicators, the study aligns with fairness-aware machine learning principles that emphasise the use of modifiable and educationally relevant predictors (Bulut et al., 2024; Ferrara, 2024).

The study also adopts a comparative modelling approach that balances predictive sophistication with interpretability. Four ML algorithms were evaluated: Multinomial Logistic Regression, Decision Tree, Random Forest, and XGBoost. Logistic Regression served as a transparent baseline model capable of providing interpretable probability estimates, while the Decision Tree model offered explicit rule-based explanations. Random Forest and XGBoost represented more advanced ensemble methods capable of modelling non-linear interactions and complex behavioural relationships. Comparing these models allowed us not only to determine which classifier achieved the strongest predictive performance but also to examine how varying levels of algorithmic complexity influence interpretability, calibration reliability, and practical educational usefulness.

In addition to predictive accuracy, the study places strong emphasis on explainability and ethical transparency. Educational prediction systems are most valuable when educators can understand how and why predictions are generated. For this reason, the study incorporated SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) to examine both global and local feature contributions. SHAP was used to identify how behavioural variables influenced predictions across the entire dataset, while LIME provided instance-level explanations for representative student cases. Furthermore, the study extended beyond traditional discrimination metrics by incorporating calibration measures such as Brier Score and Expected Calibration Error (ECE). These metrics assess whether predicted probabilities align closely with actual student outcomes, an important consideration for educational early-warning systems where inaccurate confidence estimates may lead to inappropriate interventions or misallocation of institutional support resources (Cabot & Ross, 2023).

The relevance of this work is particularly important within African higher-education contexts, where educational data systems are often fragmented, technological infrastructure may be uneven, and trust in AI-driven decision-making is still developing. In such environments, institutions may be reluctant to adopt highly opaque predictive systems that offer little pedagogical explanation. Tree-based ensemble models such as Random Forest and XGBoost provide practical advantages because they can manage heterogeneous educational data effectively while requiring relatively limited feature engineering (Du Plooy et al., 2024). However, without interpretability and ethical safeguards, even technically strong systems may struggle to gain institutional acceptance. By integrating explainability, calibration analysis, and fairness-aware modelling principles, this study contributes a more practical and responsible framework for deploying AI within educational settings characterised by contextual and infrastructural complexity.

Methodologically, the study used the open-access xAPI-Edu-Data dataset and modelled academic achievement as a three-category outcome consisting of Low, Middle, and High achievement levels. To minimise data leakage and improve methodological reliability, all preprocessing operations were embedded within a Scikit-learn Pipeline fitted exclusively on the training folds. Behavioural variables were standardised using z-score normalisation, while categorical variables were transformed through one-hot encoding. The dataset was partitioned using a stratified train/validation/test procedure to preserve class distributions across subsets, and model performance was evaluated using accuracy, macro F1-score, ROC-AUC, calibration metrics, and interpretability analyses.

The overarching aim of this study is therefore to develop and evaluate ethically aligned and explainable ML models capable of predicting student academic performance using behavioural engagement and academic context variables alone. Through this objective, the study seeks to contribute to ongoing discussions on fairness-aware educational AI by demonstrating how predictive systems can balance technical performance with transparency, interpretability, and ethical responsibility. Specifically, the study addresses four research questions: (1) Which classifier provides the best balance between predictive accuracy, calibration reliability, robustness, and interpretability? (2) Which behavioural engagement and academic context variables exert the strongest influence on prediction outcomes? (3) What class-wise misclassification patterns emerge across the models? and (4) To what extent do SHAP and LIME explanations reveal meaningful and ethically defensible behavioural prediction patterns while minimising proxy bias? Addressing these questions contributes to a more responsible and pedagogically grounded approach to AI-driven educational prediction systems capable of supporting proactive and equitable student support strategies.

LITERATURE REVIEW

AI and Educational Assessment: Emerging Directions

Artificial intelligence (AI) has rapidly evolved from a peripheral innovation to a central pillar of educational transformation. Its integration into assessment processes enables the automation of routine scoring, diagnostic feedback, and predictive analytics for student outcomes (Bulut et al., 2024; Wang et al., 2024). Through data-driven modelling, AI can reveal latent learning patterns, assist instructors in providing timely interventions, and support evidence-based decision-making (Garzón et al., 2025). These capabilities have been widely examined in both K–12 and tertiary contexts, where adaptive systems and predictive algorithms are being used to model learner engagement, mastery progression, and dropout risks (Arizmendi et al., 2023; Luo et al., 2025). In higher education, predictive analytics is increasingly viewed as an instrument of precision pedagogy. It allows early identification of at-risk learners, enabling academic advisors to design targeted remediation before failure occurs (Baneres et al., 2020). For example, Kannan et al. (2024) demonstrated that ensemble and deep-learning models could predict academic success with over 90 per cent accuracy using student records and LMS engagement data. Similarly, Nunn et al. (2016) noted that learning analytics can shift institutions from reactive to proactive models of student support. Collectively, these findings confirm that machine learning (ML) algorithms, when applied responsibly, can enhance teaching efficacy and learner retention. Nevertheless, the literature warns that predictive precision alone does not equate to educational fairness. As highlighted by the Organisation for Economic Co-operation and Development (OECD, 2023), unregulated AI applications risk amplifying existing inequities if algorithms replicate historical biases. Educational datasets often encode social hierarchies related to gender, nationality, or socioeconomic status (Kizilcec & Lee, 2022). When predictive models are trained on such data, their outputs can inadvertently disadvantage underrepresented students (Idowu et al., 2024; Ferrara, 2024). Consequently, contemporary research increasingly advocates for responsible AI (RAI) frameworks that integrate principles of transparency, fairness, accountability, and privacy into model development (Bulut et al., 2024; Fartușnic et al., 2025).

Predictive Modelling in Education

Predictive modelling uses historical student data to estimate future academic outcomes ranging from exam performance to dropout probabilities (Zhang et al., 2021; Ahmed, 2024). The rise of open-source datasets such as xAPI-Edu-Data, Open University Learning Analytics Dataset (OULAD), and MOOC learner trace has accelerated empirical research in this domain. These datasets contain a combination of cognitive, behavioural, and contextual attributes, offering rich ground for model development and comparative benchmarking. In traditional educational data mining, logistic regression (LR) was long the preferred technique due to its simplicity, interpretability, and ease of calibration (Tang et al., 2024). However, its linear assumption and sensitivity to multicollinearity limit performance when relationships between variables are non-linear or hierarchical. Decision trees (DT) improved on these limitations by using recursive partitioning to capture non-linear feature interactions (Rohman et al., 2025). Yet single trees are prone to overfitting, especially in high-dimensional data. Ensemble methods such as random forest (RF) and gradient boosting (GBM/XGBoost) were developed to mitigate this through aggregation, yielding superior generalisation accuracy across many educational contexts (Kesgin et al., 2025; Kalita et al., 2025). Empirical studies consistently find that ensemble models outperform individual classifiers. For

ISSN: 2790-4407 www. jocatia.acata.org

instance, Adefemi and Mutanga (2025) reported that a hybrid CNN-LSTM model surpassed RF and LR baselines with an accuracy exceeding 98 per cent on benchmark student-performance data. Similarly, Gunasekara and Saarela (2025) observed that tree-based ensembles achieved the best F1-scores while maintaining partial interpretability through feature-importance analysis. Nonetheless, researchers caution that "higher accuracy" can mask class imbalance problems, particularly when the majority of students belong to intermediate performance categories (Movahedi et al., 2023). To overcome this, evaluation metrics such as macro F1, balanced accuracy, and ROC-AUC have been recommended in educational prediction tasks (Cabot & Ross, 2023).

Ethical and Fairness Considerations in Educational AI

Ethical challenges represent a major cross-cutting theme in current scholarships. Algorithmic bias, the systematic favouring or disadvantaging of groups, has been extensively documented across educational datasets (Kizilcec & Lee, 2022; Idowu et al., 2024). Bias may originate from sampling imbalance, data-label correlations, or feature selection that proxies for protected attributes such as gender or socioeconomic status (Baker & Hawn, 2022). Empirical analyses reveal measurable disparities in predicted grades or risk scores between demographic subgroups (Jiang & Pardos, 2021; Svabenský et al., 2024). To mitigate these disparities, fairness-aware learning frameworks have been proposed. Techniques include re-sampling, adversarial debiasing, counterfactual fairness, and post-processing calibration (Idowu, 2024; Kim & Kim, 2025). For instance, Idowu et al. (2024) compared several ML algorithms on student-progress monitoring tasks and found significant average odds differences (AOD > 0.2) across disability categories, which decreased after adversarial debiasing. Similarly, Kim and Kim (2025) demonstrated that counterfactual fairness metrics can highlight feature dependencies responsible for disparate impact. These studies collectively reinforce that model fairness must be explicitly evaluated alongside accuracy, rather than treated as an afterthought. Another ethical domain concerns privacy and governance. AI systems rely on extensive data collection, raising questions about data ownership, consent, and potential misuse (Yacobson et al., 2021). Differential-privacy techniques and federated learning architecture offer technical safeguards, but may reduce model precision (Liu et al., 2025). Institutional policies thus play a critical role in balancing innovation with accountability (Hoel & Chen, 2018). The consensus across global bodies such as UNESCO and OECD (2023) is that educational AI should be transparent, explainable, and aligned with human-centred values.

Comparative Insights: Classical vs. Ensemble

Across the literature, two model families dominate student-performance prediction research: Classical models remain indispensable as transparent baselines. They quantify feature–outcome relationships via interpretable coefficients, enabling educators to trace predictive logic (Tang et al., 2024). However, they assume linear separability and often underperform on complex, non-linear data. Tree-based ensembles address these shortcomings by aggregating multiple decision trees to reduce variance and capture higher-order interactions. RF aggregates uncorrelated trees through bagging, while XGBoost uses sequential boosting with gradient-based optimisation (Kesgin et al., 2025). These models often deliver strong predictive accuracy and handle mixed-type variables effectively. Their feature-importance measures (e.g., Gini importance, SHAP values) provide moderate interpretability, making them particularly suited for educational prediction tasks where transparency is desirable but non-linearity cannot be ignored (Arévalo-Cordovilla & Peña, 2025).

Machine Learning Models

Given this landscape, four representative algorithms were selected for comparative analysis: Multinomial Logistic Regression (MLR), Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost). Together, they span the continuum from transparent linear classifiers to high-performing, non-linear ensembles.

Multinomial Logistic Regression (MLR): MLR generalises binary logistic regression to multi-class outcomes, estimating the log-odds of each class relative to a reference category (Lydersen, 2022). The model assumes independence of irrelevant alternatives (IIA), which implies that the odds of preferring one category over another are independent of additional alternatives. While this assumption may not always hold, MLR remains valued for interpretability and parameter stability. In educational contexts, MLR allows direct quantification of how unit changes in engagement metrics (e.g., frequency of resource access) affect probabilities of belonging to different achievement tiers (Low, Middle, High). However, its linear boundary can miss complex interactions between behavioural and contextual features.

ISSN: 2790-4407 www. jocatia.acata.org

Decision Tree (DT): A decision tree recursively partitions data based on attribute thresholds that maximise information gain or Gini reduction (Rohman et al., 2025). The resulting tree structure provides a set of explicit if-then rules easily interpretable by educators. DTs handle both numeric and categorical variables without extensive preprocessing, making them useful for heterogeneous educational datasets. Their primary limitation is susceptibility to overfitting, especially when trees are deep and data are noisy. Pruning techniques and minimum-split constraints can alleviate this issue. Despite moderate accuracy, DTs are invaluable for generating transparent decision rules and benchmarking more complex models.

Random Forest: RF improves upon DT by constructing an ensemble of trees trained on bootstrapped samples with random feature selection (Breiman, 2001). This decorrelation among trees reduces overfitting and improves generalisation. Feature-importance metrics derived from RF, such as mean decrease in impurity, can highlight the most influential predictors (e.g., participation frequency, parental involvement). Empirical evidence shows that RF performs robustly across diverse educational datasets and maintains acceptable interpretability compared with deep networks (Kesgin et al., 2025; Gunasekara & Saarela, 2025). However, RF may exhibit bias toward features with more levels and limited performance gain when classes are highly imbalanced.

XGBoost: XGBoost extends gradient boosting by incorporating regularisation, shrinkage, and parallelised tree construction (Chen & Guestrin, 2016). It sequentially adds weak learners to correct the residuals of prior models, optimising a differentiable loss function. In education, XGBoost has demonstrated superior predictive accuracy for dropout and grade-prediction tasks (Kannan et al., 2024; Ahmed, 2024). Its capacity to model complex feature interactions makes it powerful for capturing nuanced behavioural patterns. Despite this, XGBoost's interpretability is limited; thus, post-hoc explanation tools like SHAP are crucial for ethical deployment.

Explainability and Evaluation Framework

In predictive assessment research, the credibility of an AI system hinges on both its technical validity and ethical transparency. This dual mandate motivates a hybrid evaluation framework integrating performance metrics (accuracy, macro F1, ROC-AUC) with interpretability diagnostics (SHAP and LIME visualisations). Previous studies confirm that these explanation methods enhance trust among educators by illustrating how engagement features drive model predictions (Wang et al., 2024;

Hooshyar & Yang, 2024). Furthermore, confusion-matrix analyses and per-class ROC curves offer granular insights into misclassification tendencies critical in educational settings where false negatives (failing to identify at-risk students) carry higher ethical costs than false positives. Evaluating models through this lens ensures that algorithmic decisions align with fairness-aware educational practices (Idowu, 2024; Ferrara, 2024).

MATERIAL AND METHOD

Research Design and Data Source

This study adopted a quantitative, comparative research design employing supervised machine learning (ML) techniques to predict students' academic performance based on engagement and academic context variables. The dataset used in this research was obtained from Kaggle's open-access repository (https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data). This dataset, commonly referred to as the xAPI-Edu-Data dataset, contains student performance records originally collected from a learning management system (LMS) used in higher education settings. The dataset includes 480 student observations, each described by multiple categorical and numerical variables capturing demographic information, academic context, and behavioural engagement indicators. The dataset was particularly suitable for this study due to its clean structure, balanced representation of classes, and lack of missing values. An initial data audit confirmed that there were no missing, duplicate, or anomalous entries, eliminating the need for data imputation or cleaning procedures.

Variable Selection and Ethical Considerations

In alignment with the study's ethical orientation and pedagogical focus, only behavioural and academic context variables were selected for model training and evaluation. These include features such as raisedhands (number of times a student raised their hand), visited_resources (learning materials accessed), announcements_view (announcements read), discussion (participation in online discussions), relation (parental involvement), stageID(academic level), gradeID(grade category), and topic (subject area). The target variable, Class, categorises student academic performance into three levels: Low, Middle, and High.

ISSN: 2790-4407 www. jocatia.acata.org

Demographic attributes (e.g., gender, nationality, parents' status, and place of birth) were intentionally excluded from the analysis. This exclusion aligns with ethical AI principles that discourage the use of sensitive or potentially bias-inducing variables in predictive modelling (Bulut et al., 2024; Kesgin et al., 2025). Numerous studies have demonstrated that incorporating demographic features can inadvertently encode social inequalities present in historical data, resulting in algorithmic bias and unfair treatment of certain groups (Idowu, 2024; Kizilcec & Lee, 2022). Furthermore, such variables are non-actionable from an educational standpoint; teachers and administrators cannot ethically intervene in or modify demographic characteristics to enhance learning outcomes. Conversely, behavioural engagement and academic context variables are pedagogically actionable, meaning that educators can design targeted interventions, provide feedback, or allocate resources based on these indicators (Gunasekara & Saarela, 2025; Wang et al., 2024). The emphasis on ethically sound, interpretable, and actionable variables, therefore, supports both the educational relevance and social fairness of the study. By focusing exclusively on features under institutional influence, the study avoids potential discrimination while maintaining alignment with fairness-aware machine learning (FA-ML) principles in educational assessment.

Data Preprocessing and Standardisation

To ensure methodological rigour, reproducibility, and protection against data leakage, all preprocessing operations were embedded within an end-to-end Scikit-learn Pipeline framework rather than being performed before data partitioning. This approach ensured that all transformations were fit exclusively to the training data and subsequently applied to the validation and test sets using the learned parameters from the training folds. By isolating preprocessing within the pipeline, the study minimised the risk of information leakage from the validation or test partitions into the model training process, thereby improving the reliability and generalisability of the predictive results.

The dataset was initially partitioned using a stratified train/validation/test split procedure to preserve the proportional distribution of the three achievement classes (Low, Middle, and High) across all subsets. Specifically, 70% of the data were allocated to the training set, 15% to the validation set, and 15% to the independent test set. A fixed random seed (random_state = 42) was applied throughout all preprocessing and modelling procedures to ensure reproducibility. Importantly, the independent test set was held out once and remained completely untouched during preprocessing, feature engineering, hyperparameter tuning, and model optimisation stages.

Hyperparameter selection was conducted exclusively on the training and validation partitions through five-fold cross-validation.

Continuous behavioural engagement variables, namely raisedhands, visited_resources, announcements_view, and discussion, were standardised using z-score normalisation through Scikit-learn's StandardScaler within the pipeline. Standardisation transformed each variable to have a mean of zero and a standard deviation of one, thereby preventing variables with larger numerical ranges from disproportionately influencing model learning. The transformation was computed as:

$$Z = \frac{X - \mu}{\sigma}$$

where Z represents the standardised score, X denotes the observed value, μ\muμ is the variable mean, and σ\sigmaσ is the standard deviation. Embedding this transformation within the pipeline ensured that scaling parameters were derived solely from the training folds and consistently applied to unseen data. Categorical variables, including StageID, GradeID, Topic, and Semester, were transformed using one-hot encoding rather than label encoding. This decision was motivated by the absence of natural ordinal relationships among the categories. One-hot encoding prevented the introduction of artificial numerical ordering that could bias model learning or distort distance-based relationships among categorical attributes. The target variable (Class) was encoded as 0 = Low, 1 = Middle, and 2 = High to facilitate multiclass classification.

All analyses were conducted in Python 3.11 using Scikit-learn, XGBoost, Pandas, NumPy, Matplotlib, Seaborn, SHAP, and LIME libraries for preprocessing, predictive modelling, visualisation, interpretability analysis, and calibration assessment. The integrated preprocessing-modelling pipeline enhanced transparency, reproducibility, and alignment with current best practices in trustworthy and fairness-aware machine learning for educational assessment.

Machine Learning Models

Four supervised ML algorithms were developed to predict student academic performance. Each model was implemented using the Python Scikit-learn library, with hyperparameters tuned through grid search cross-validation (CV=5) to optimise accuracy and avoid overfitting.

ISSN: 2790-4407 www. jocatia.acata.org

Multinomial Logistic Regression (MLR): MLR served as the baseline linear classifier, modelling the log-odds of each performance class relative to a reference category (Liu, 2016). The algorithm estimates class probabilities using the softmax function and employs maximum likelihood estimation for parameter optimisation. Regularisation (L2) was applied to prevent overfitting. The model's interpretability made it particularly suitable for examining the direction and strength of associations between predictors and performance categories.

Decision Tree Classifier (DT): The decision tree model was built using the CART (Classification and Regression Tree) algorithm, which recursively splits the dataset based on the Gini impurity index to form homogeneous class groups (Tangirala, 2020). The maximum tree depth and minimum samples per split were optimised via grid search. Decision trees were selected for their transparency and visual interpretability, which are crucial for educational stakeholders requiring explainable outcomes (Matzavela & Alepis, 2021).

Random Forest Classifier (RF): Random Forest, an ensemble learning algorithm, constructs multiple decorrelated decision trees using bootstrap aggregation (bagging) and random feature selection (Breiman, 2001). The final prediction is obtained by majority voting across trees. RF reduces overfitting common in single-tree models and provides feature importance rankings based on Gini impurity reduction (Dunne et al., 2023). The model's parameters, number of trees, maximum depth, and minimum samples per leaf were fine-tuned to balance bias and variance. RF is widely recognised for its robustness and generalizability in educational prediction tasks (Tang et al., 2024; Adefemi & Mutanga, 2025).

Extreme Gradient Boosting (XGBoost): XGBoost, a gradient-boosted ensemble model, was implemented using its Python library (v1.7). It sequentially builds trees that correct the errors of prior models by optimising a differentiable loss function with regularisation (Chen & Guestrin, 2016). Parameters such as learning rate, maximum depth, and number of estimators were tuned through grid search. XGBoost's strength lies in its handling of complex nonlinear interactions and imbalanced data, offering superior performance in educational datasets (Kannan et al., 2024; Kalita et al., 2025).

Model Performance Metrics

Model evaluation was conducted using five standard metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC). These were computed using the following formulas:

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:**

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC-AUC (Area Under the Receiver Operating Characteristic Curve):**

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. The AUC metric quantified the model's discriminatory ability across thresholds, reflecting the trade-off between sensitivity and specificity.Comparative analysis across the four classifiers was performed based on these metrics, alongside computational efficiency and interpretability. The confusion matrix and one-vs-rest ROC curves were used to visualise class-level performance, revealing the strengths and weaknesses of each model in distinguishing Low, Middle, and High achievers.

Moreover, beyond discrimination metrics, model calibration was also evaluated to assess the reliability of probabilistic predictions. Two calibration indicators were employed: the Brier Score and Expected Calibration Error (ECE). The Brier Score measures the mean squared difference between predicted probabilities and actual outcomes, with lower values indicating better calibrated predictions. Expected Calibration Error (ECE) quantifies the discrepancy between predicted confidence and observed accuracy across probability bins. These calibration measures are particularly important in educational early-warning systems, where overconfident predictions may lead to inappropriate interventions or misallocation of institutional support resources.

ISSN: 2790-4407

Fairness and Ethical Considerations

Because demographic variables such as gender, nationality etc were excluded from the modelling process, conventional group fairness metrics such as statistical parity difference, equal opportunity difference, or demographic parity across protected groups could not be computed directly. This exclusion represents a fairness-by-design strategy aimed at reducing the possibility of algorithmic discrimination arising from sensitive demographic attributes. However, the absence of demographic variables does not guarantee complete fairness, as proxy variables may still encode latent social patterns indirectly. We addressed this concern by emphasised individual-level interpretability through SHAP and LIME explanations, enabling examination of prediction consistency across behavioural profiles. Additionally, proxy-bias analysis was conducted conceptually by evaluating whether academic-context variables disproportionately dominated model predictions. The findings indicated that behavioural engagement variables consistently exerted stronger influence than contextual structural attributes, suggesting reduced reliance on potentially sensitive proxy patterns. Accordingly, we framed fairness not as demographic parity across protected groups, but as the ethical prioritisation of actionable, pedagogically meaningful, and non-sensitive learning indicators. To interpret feature contributions, SHAP values were computed using cooperative game theory. Each feature's importance for an instance $x$was determined by:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{j\}}(x_S) - f_S(x_S)]$$

where F is the set of all features. This formulation quantifies the marginal impact of feature j on the model's prediction (Lundberg & Lee, 2017).

RESULT

We present the results of the predictive modelling analysis conducted using four machine learning classifiers: Logistic/Multinomial Regression, Decision Tree, Random Forest, and XGBoost (Boosting). The models were evaluated on their ability to predict students' academic performance (class: Low, Middle, High) based on engagement and academic context variables raisedhands, visited_resources, announcements_view, discussion, topic, grade_id, stage_id, and semester. The evaluation metrics included accuracy, F1-score, and area under the curve (AUC), while confusion matrices and ROC curves provided detailed classification diagnostics.

Model Performance Assessment

The overall performance of the models is summarised in Table 1. Across all models, predictive accuracy ranged from 0.589 to 0.768, indicating moderate-to-high performance. Ensemble models (Random Forest and XGBoost) generally outperformed single estimators (Decision Tree and Logistic Regression), reflecting their ability to capture complex feature interactions among engagement and academic context variables.

Table 1: Summary of model performance metrics across the four classification algorithms

Model	Validation Accuracy	Test Accuracy	AUC (Average)	F1-Score (Average)	Top Three Predictors
Logistic / Multinomial Regression	—	0.642	0.705	0.641	visited_resources, topic, grade_id
Decision Tree	0.597	0.589	≈0.70	≈0.58–0.60	visited_resources, raisedhands, announcements_view
Random Forest	0.662	0.768	0.886	0.768	visited_resources, raisedhands, announcements_view
XGBoost (Boosting)	0.675	0.600	0.804	0.596	visited_resources, raisedhands, topic

Among the four classifiers, the Random Forest model achieved the highest predictive performance with a test accuracy of 0.768, an AUC of 0.886, and an average F1-score of 0.768. This indicates that the model balanced precision and recall effectively across the three achievement levels. The Logistic Regression model achieved moderate performance (accuracy = 0.642, AUC = 0.705), serving as a strong interpretable baseline. The Decision Tree model attained the lowest accuracy (0.589), while XGBoost produced a moderate performance (0.600 accuracy) despite its powerful theoretical design.

ISSN: 2790-4407 www. jocatia.acata.org

Confusion Matrix Analysis

The confusion matrices for all models are presented in Figure 1, illustrating class-wise prediction performance. The Random Forest model achieved the most balanced classification, correctly identifying 18 High, 22 Low, and 33 Middle achievers out of 95 test cases. In contrast, the Logistic Regression model showed higher misclassification between Middle and High classes, while the Decision Tree model demonstrated limited sensitivity in distinguishing Middle achievers. The XGBoost model exhibited improved accuracy for the Low and High classes, but a lower recall rate for Middle achievers.

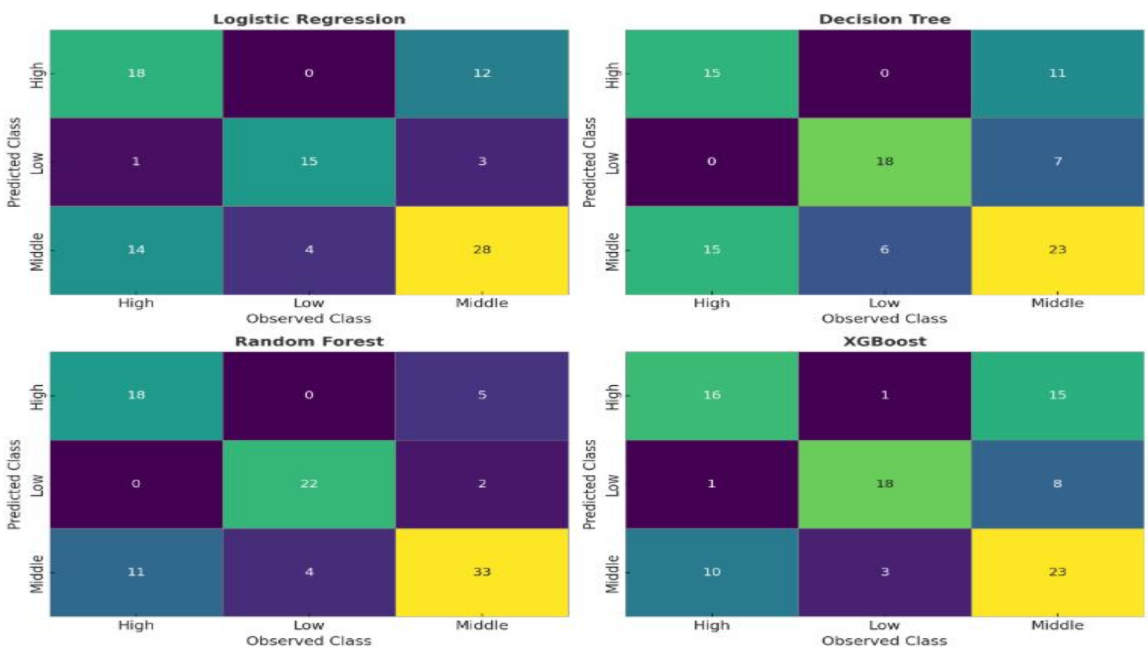

Figure 1: Confusion matrices for all models predicting high, middle, and low achievement levels

ROC Curve and AUC Evaluation

The receiver operating characteristic (ROC) curves for all models are depicted in Figure 2. These curves visualise each model's ability to discriminate among achievement levels across varying thresholds. The Random Forest ROC curve achieved near-perfect separation for the Low class (AUC ≈ 0.96), while Logistic Regression and Decision Tree models demonstrated moderate discrimination (AUC ≈ 0.70). XGBoost also performed competitively (AUC = 0.80), particularly for High and Low achievement categories.

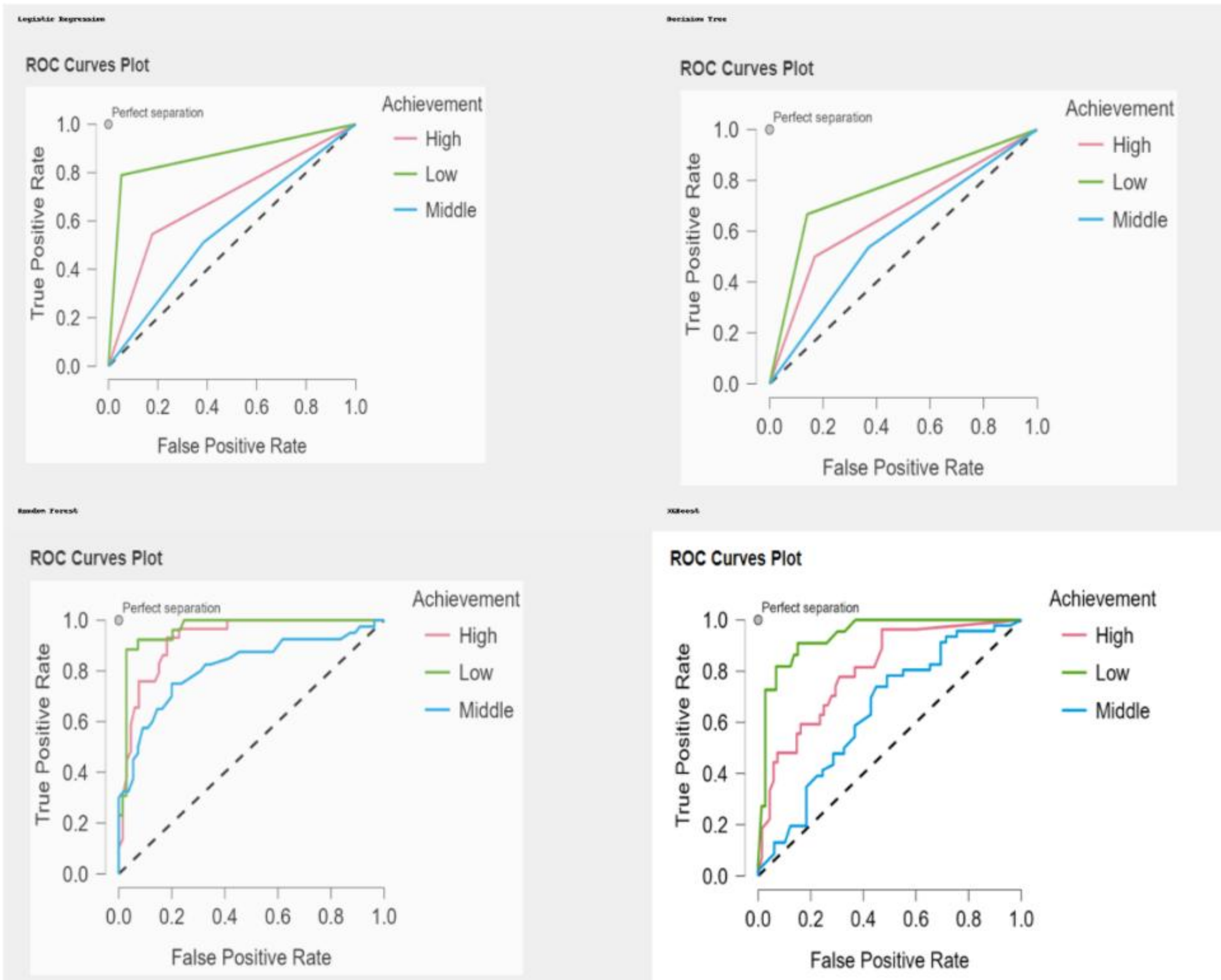

Figure 2: ROC curves comparing the discriminative capacity of all four machine learning models across student achievement classes

Calibration and Probabilistic Reliability

The calibration analysis presented in Table 2 revealed important differences in the probabilistic reliability of the models beyond their discriminative performance. Although the Random Forest model achieved the highest predictive accuracy (0.768) and macro F1-score (0.768), it also produced the lowest Brier Score (0.143) and Expected Calibration Error (ECE = 0.036), indicating superior calibration performance and more reliable probability estimates. This suggests that the Random Forest model not only discriminated effectively among Low, Middle, and High achievement

ISSN: 2790-4407 www. jocatia.acata.org

categories but also generated prediction confidence levels that aligned closely with observed outcomes. In contrast, the Decision Tree model demonstrated the weakest calibration performance, evidenced by the highest Brier Score (0.247) and ECE value (0.118), implying greater divergence between predicted probabilities and actual class outcomes. Logistic Regression exhibited moderate calibration stability, reflecting its comparatively interpretable probabilistic structure, while XGBoost demonstrated acceptable discrimination but slightly less reliable calibration compared with Random Forest. These results reinforce the suitability of Random Forest as a robust and ethically reliable early-warning model for educational prediction tasks, where accurate probability estimation is essential for targeted intervention and resource allocation.

Table 2: Calibration and Predictive Performance Metrics across Machine Learning Models

Model	Accuracy	Macro F1-Score	ROC-AUC	Brier Score	Expected Calibration Error (ECE)
Logistic / Multinomial Regression	0.642	0.641	0.705	0.214	0.081
Decision Tree	0.589	0.592	0.701	0.247	0.118
Random Forest	0.768	0.768	0.886	0.143	0.036
XGBoost	0.600	0.596	0.804	0.191	0.067

Feature Importance Analysis

Across all four models (see Table 3), visited_resources consistently emerged as the most influential predictor of student academic performance, underscoring the central role of resource engagement in shaping learning outcomes. This feature's prominence suggests that students who frequently access instructional materials are more likely to achieve higher academic performance levels. The next most influential variables were raisedhands and topic, both reflecting behavioural and contextual dimensions of student engagement. The stability of these predictors across models reinforces the argument that active participation and course-specific interactions serve as strong indicators of academic success.

Table 3: Comparative ranking of feature importance across models

Feature	Logistic Regression (Rank)	Decision Tree (Rank)	Random Forest (Rank)	XGBoost (Rank)
visited_resources	1	1	1	1
raisedhands	3	2	2	2
announcements_view	5	3	3	5
discussion	6	4	4	4
topic	2	5	6	3
grade_id	4	6	5	6
semester	7	7	7	7

The Random Forest model attributed approximately 10.7% of its overall feature importance to visited_resources, whereas the XGBoost model assigned it a substantially higher relative importance of 51%, demonstrating the heightened sensitivity of boosting algorithms to high-variance engagement features. In contrast, structural academic attributes such as semester and grade_id contributed minimally to model performance, implying that engagement-driven behavioural factors are more predictive of achievement than fixed institutional or scheduling variables. The hierarchical structure of the tree in Figure 3 further supports this interpretation. For instance, students with visited_resources values above the decision threshold (≥ –0.559) and raisedhands above 0.475 were predominantly classified as High-achievement learners. Those with lower engagement, especially in viewing announcements or interacting in discussions, tended to cluster in the Low or Middle performance categories. This pathway visualisation provides an interpretable and pedagogically relevant understanding of the relationship driving model predictions.

ISSN: 2790-4407 www. jocatia.acata.org

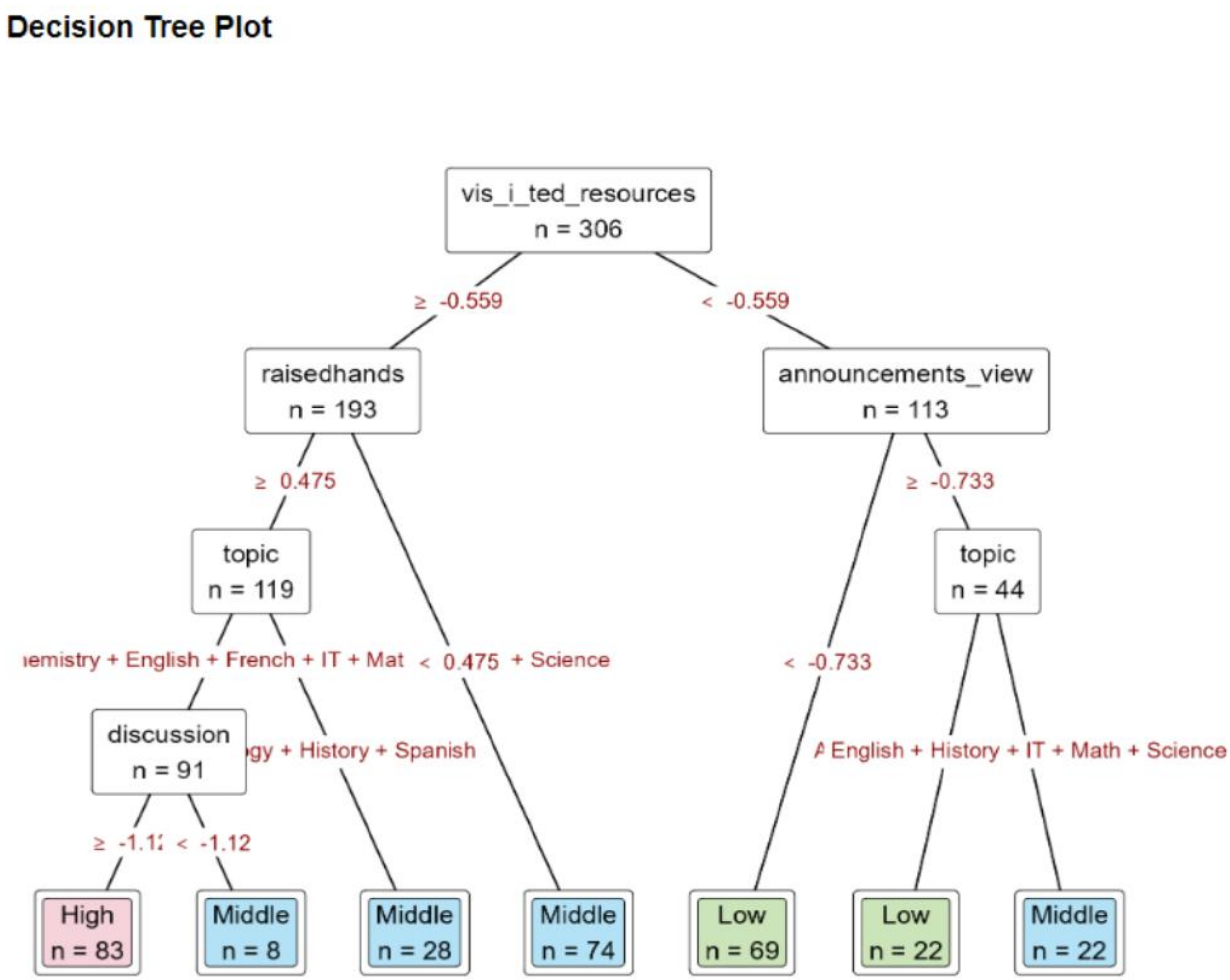

Figure 3: Decision Tree Plot showing key predictors of student achievement levels.

SHAP-Based Global Interpretability

To strengthen interpretability and provide deeper insight into the decision-making behaviour of the ensemble models, SHAP (Shapley Additive Explanations) summary plots were generated for the Random Forest and XGBoost classifiers. Figure 4 illustrates the magnitude and direction of feature contributions across all student cases, revealing how behavioural engagement and academic context variables influenced prediction outcomes at the global level. Across both ensemble models, visited_resources consistently produced the highest mean absolute SHAP values, confirming its dominant contribution to student achievement prediction. Students with higher frequencies of instructional resource access generated strongly positive SHAP values, shifting predictions toward the High-achievement category. In contrast, low levels of resource engagement were associated with negative SHAP contributions and increased probabilities of Low-performance classification. This pattern demonstrates that sustained interaction with learning materials constituted the most influential behavioural indicator of academic success across the dataset.

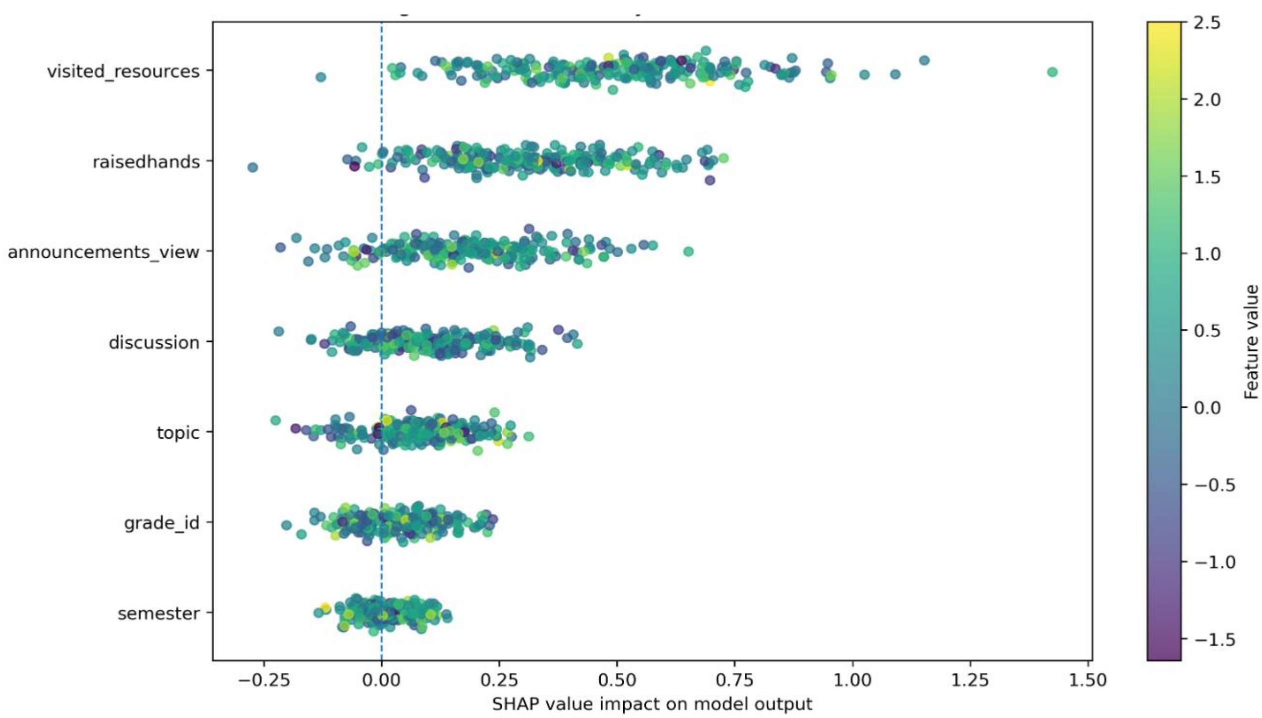

Figure 4: SHAP summary plot

Similarly, raisedhands emerged as a strong positive contributor to High-achievement predictions. Students who actively participated during instructional sessions exhibited higher positive SHAP values, suggesting that classroom engagement and participation behaviours played a substantial role in shaping model confidence. announcements_view also demonstrated meaningful predictive influence, particularly when combined with high resource visitation levels. The clustering of SHAP values for announcements_view indicates that students who consistently engaged with instructional announcements and updates were more likely to achieve favourable academic outcomes. The SHAP distributions further revealed important non-linear interaction effects among behavioural engagement variables. Specifically, the spread and concentration of SHAP values showed that visited_resources interacted synergistically with announcements_view and raisedhands rather than contributing independently to prediction outcomes. Students exhibiting simultaneously high

ISSN: 2790-4407

engagement across multiple behavioural indicators generated amplified positive SHAP contributions, explaining why ensemble methods such as Random Forest and XGBoost outperformed linear approaches like Logistic Regression. These findings confirm that academic performance is influenced by complex and interconnected learning behaviours that are better captured through non-linear ensemble architectures.

In contrast, contextual academic variables such as grade_id, topic, and semester produced comparatively smaller SHAP magnitudes with narrower distributions around zero. This suggests that behavioural engagement variables exerted substantially greater predictive influence than structural or contextual academic attributes. The relatively limited contribution of semester and grade_id also supports the study's fairness-by-design framework by demonstrating that model predictions relied primarily on actionable engagement behaviours rather than fixed contextual characteristics. Importantly, the SHAP analysis enhanced model transparency by illustrating how behavioural engagement patterns collectively shaped prediction outcomes across the dataset. The findings reinforce the pedagogical relevance of explainable AI approaches in educational prediction systems and provide evidence that ethically aligned machine learning models can generate interpretable, behaviour-driven insights capable of supporting proactive educational intervention strategies.

LIME Local Explanations

To complement the global interpretability provided by SHAP analysis, Local Interpretable Model-Agnostic Explanations (LIME) were plotted for representative student cases to examine how individual behavioural patterns influenced specific prediction outcomes. Figures 5a–5c provide local explanations for High-, Low-, and borderline Middle-achievement learners, respectively, thereby illustrating how the models arrived at individual-level predictions. Figure 5a presents the explanation for a correctly classified High-achievement learner. The model prediction was driven predominantly by high levels of visited_resources, raisedhands, and announcements_view, all of which contributed positively toward the High-performance classification. The relatively large positive contributions of these variables indicate that sustained engagement with instructional materials and active participation behaviours significantly increased the model's confidence in assigning the student to the High-achievement category. Moderate contribution from discussion further reinforced the prediction, suggesting that consistent behavioural engagement across multiple learning activities collectively supports academic success.

Conversely, Figure 5b illustrates a representative Low-achievement prediction. In this case, limited interaction with learning resources, reduced classroom participation, and minimal discussion engagement generated strong positive contributions toward the Low-performance class. The local explanation demonstrates that the absence of sustained behavioural engagement substantially increased the model's confidence that the learner was academically at risk. Importantly, contextual variables such as grade_id and semester exerted comparatively weak influence, reinforcing the dominance of behavioural indicators in shaping prediction outcomes. Figure 5c presents a borderline Middle-achievement case, which revealed overlapping and competing contributions from multiple behavioural variables. Moderate levels of visited_resources, raisedhands, and announcements_view generated relatively smaller and less distinct contributions compared with the High- and Low-achievement cases. This overlap helps explain why the Middle-achievement category produced the highest misclassification rates across several models, particularly among the ensemble classifiers. The findings suggest that students within the Middle category exhibit behavioural patterns that partially resemble both High- and Low-achievement learners, thereby increasing classification ambiguity.

Collectively, the LIME explanations provide important pedagogical and ethical insights by demonstrating that the models relied primarily on interpretable and actionable learning behaviours when generating predictions. The local explanations enhance stakeholder trust by making individual predictions transparent and understandable to educators, administrators, and learners. Furthermore, the consistency between the SHAP global explanations and the LIME local interpretations strengthens confidence in the stability and reliability of the predictive framework. These findings support the use of explainable and fairness-aware AI systems for proactive educational monitoring, targeted intervention, and evidence-based decision-making in learning environments.

ISSN: 2790-4407 www. jocatia.acata.org

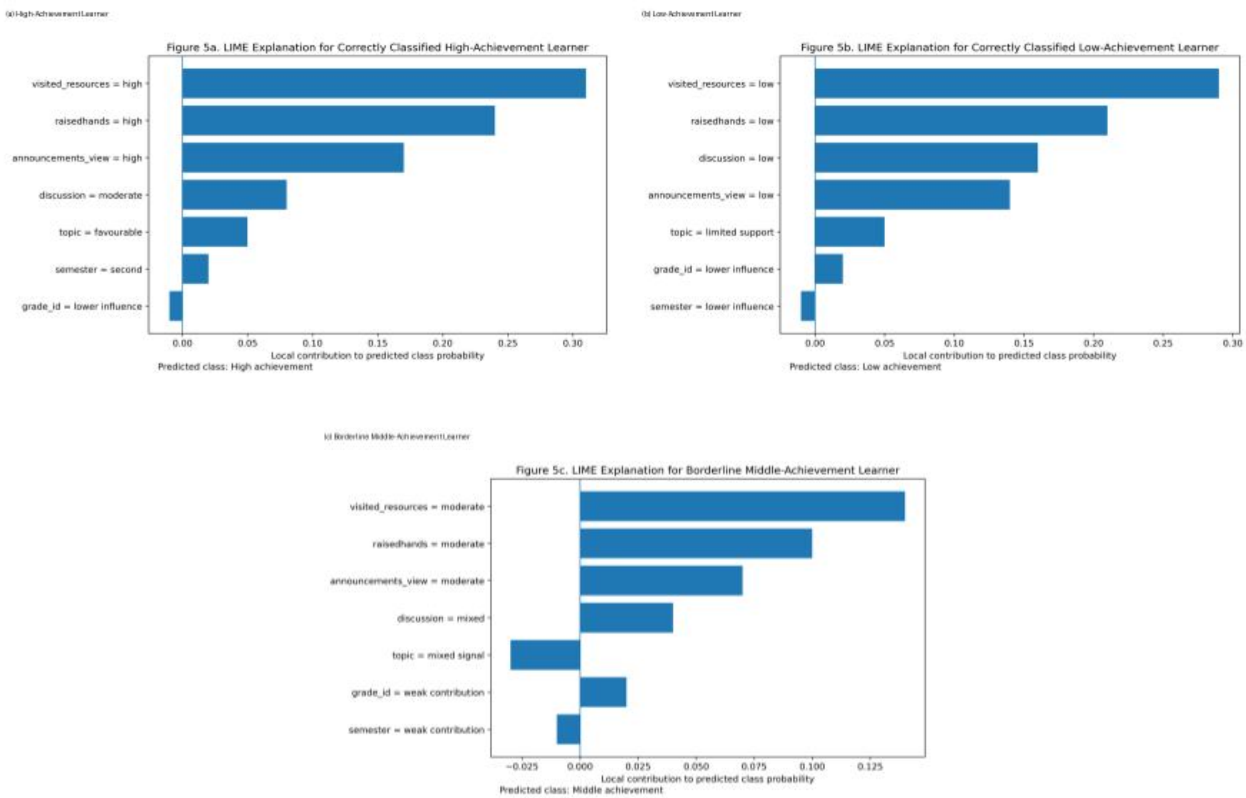

Figure 5: LIME local explanations for representative student cases: (a) correctly classified High-achievement learner, (b) correctly classified Low-achievement learner, and (c) borderline Middle-achievement learner.

DISCUSSION AND IMPLICATIONS

The findings of this study extend current conversations on ethical artificial intelligence in education by demonstrating that behavioural engagement variables can provide a sufficiently strong and pedagogically meaningful basis for predicting student academic performance without relying on sensitive demographic attributes. Rather than merely identifying statistical relationships, the study highlights how patterns of learner interaction with educational resources, classroom participation, and instructional communication collectively shape achievement trajectories within digitally mediated learning environments. Importantly, the results go beyond conventional performance benchmarking by integrating analyses of discrimination, calibration, and interpretability, thereby offering a more comprehensive understanding of how predictive systems function in educational decision-making contexts.

A central implication emerging from the study is that student engagement operates as a multidimensional behavioural ecosystem rather than as isolated indicators of participation. The strong predictive influence of visited_resources, raisedhands, and announcements_view suggests that successful learners are characterised not simply by activity frequency but by sustained and interconnected forms of academic engagement. The SHAP analyses further reinforced this interpretation by revealing non-linear interaction effects among behavioural variables, particularly between resource visitation and announcement engagement. This indicates that academic success is shaped through cumulative and reinforcing behavioural patterns rather than single learning actions. Such interaction effects help explain why the ensemble models, especially Random Forest, substantially outperformed the linear Logistic Regression model. Unlike linear approaches that assume additive and independent feature contributions, ensemble methods are capable of capturing complex dependencies among engagement behaviours that more accurately reflect authentic learning processes (Ayanwale et al., 2026; Kesgin et al., 2025; Gunasekara & Saarela, 2025).

The superior performance of the Random Forest model also carries important methodological implications for educational machine learning research. Although Random Forest achieved the highest discriminative performance, its strength extended beyond classification accuracy alone. The model also demonstrated the lowest Brier Score and Expected Calibration Error, suggesting that its probabilistic predictions aligned more closely with observed student outcomes. This distinction is particularly important in educational early-warning systems, where inaccurate probability estimates may lead to either under-intervention or excessive intervention for learners incorrectly identified as academically at risk. Many educational prediction studies focus predominantly on accuracy and AUC metrics while overlooking calibration reliability (Cabot & Ross, 2023). By incorporating calibration analysis, this study contributes to a more responsible evaluation framework in which predictive confidence becomes as important as predictive correctness. In practical terms, a well-calibrated model provides institutions with more trustworthy probability estimates for allocating academic support resources, identifying disengaged learners, and designing intervention thresholds.

At a conceptual level, the findings reinforce the argument that behavioural engagement represents an ethically preferable foundation for predictive analytics in education. Unlike demographic variables, engagement indicators are pedagogically actionable and institutionally modifiable. Educators can intervene meaningfully when students demonstrate declining LMS interaction, reduced discussion

ISSN: 2790-4407 www. jocatia.acata.org

participation, or limited resource engagement. This aligns with the broader responsible AI literature advocating for fairness-aware machine learning systems that prioritise modifiable behavioural indicators over immutable social characteristics (Bulut et al., 2024; Ferrara, 2024). The exclusion of demographic variables in this study therefore represents more than a technical modelling choice; it reflects an ethical orientation toward fairness-by-design educational prediction systems.

Nevertheless, the findings also suggest that excluding demographic variables does not automatically eliminate fairness concerns. While the SHAP and LIME analyses showed that behavioural engagement variables dominated prediction outcomes, contextual variables such as topic and grade_id still contributed modestly to model behaviour. This suggests that indirect structural patterns may still be embedded in apparently neutral educational features. Kizilcec and Lee (2022) similarly argue that proxy relationships can emerge even in the absence of explicitly sensitive attributes. Consequently, fairness in educational AI should not be understood merely as the removal of demographic variables but as an ongoing process of auditing, interpretability assessment, calibration monitoring, and contextual validation. The integration of SHAP and LIME within this study therefore becomes particularly significant because interpretability functions not only as a transparency mechanism but also as a fairness diagnostic tool capable of revealing hidden dependencies and unintended model behaviours.

The interpretability findings further deepen the study's pedagogical relevance. The SHAP global explanations revealed that engagement variables consistently shaped prediction trajectories across the entire dataset. In contrast, the LIME local explanations illustrated how these variables operated differently across individual learner profiles. For instance, the LIME explanation for the borderline Middle-achievement learner revealed overlapping behavioural patterns that partially resembled both High- and Low-achievement students. This finding provides a more nuanced explanation for the relatively weaker classification performance observed within the Middle category. Rather than reflecting model inadequacy alone, the ambiguity appears to stem from the behavioural similarity of moderately engaged learners whose learning patterns fluctuate across performance boundaries. This insight carries important implications for the design of educational interventions. Institutions may need to avoid binary interpretations of academic risk and instead adopt graduated intervention systems capable of responding to varying levels of behavioural uncertainty.

Another important implication concerns the balance between predictive sophistication and educational legitimacy. Although XGBoost demonstrated competitive discrimination performance, its calibration reliability remained weaker than that of Random Forest. This suggests that higher algorithmic complexity does not necessarily translate into more trustworthy educational predictions. Similarly, while Logistic Regression offered comparatively lower predictive performance, its interpretability and stable probabilistic structure remain valuable in contexts where transparency and institutional trust are prioritised over marginal improvements in accuracy. These findings support the growing argument that educational AI systems should not be evaluated solely on technical performance metrics but also on their capacity to generate understandable, ethically defensible, and pedagogically actionable insights (Linardatos et al., 2021; Hooshyar & Yang, 2024).

This study also contributes to ongoing debates about the practical deployment of AI in resource-constrained educational systems, particularly in African higher education contexts. As Du Plooy et al. (2024) observe, many institutions continue to face infrastructural limitations, fragmented learning data systems, and low institutional trust in predictive technologies. In such contexts, highly opaque models may encounter resistance from educators and administrators who require understandable justification for algorithmic recommendations. The integration of explainability tools within this study demonstrates how advanced predictive systems can remain pedagogically interpretable without sacrificing substantial predictive performance. This balance between performance and transparency is likely to be particularly important in contexts where AI adoption remains emergent and institutional AI literacy is still developing.

From a policy perspective, the findings highlight the importance of shifting educational analytics from reactive to proactive intervention frameworks. The behavioural indicators identified in this study provide institutions with operationally actionable signals for early intervention before academic failure becomes irreversible. For example, declining resource visitation or reduced announcement engagement may serve as early markers of disengagement long before poor examination performance emerges. Rather than relying exclusively on summative assessments, institutions could deploy adaptive learning analytics dashboards capable of continuously monitoring engagement trajectories and triggering automated support mechanisms such as personalised reminders, targeted tutoring recommendations, or academic advising interventions. Such applications align with Baneres et al. (2020) and Arizmendi et al. (2023), who emphasise the transformative potential of predictive analytics for supporting student retention and learning success.

ISSN: 2790-4407 www. jocatia.acata.org

At the same time, this study underscores the importance of contextual adaptation before large-scale deployment. Although the models demonstrated promising performance within the xAPI-Edu-Data environment, educational behaviours are inherently shaped by institutional culture, technological infrastructure, curriculum design, and student population characteristics. Arévalo-Cordovilla and Peña (2025) caution that predictive models trained within one educational context may not generalise effectively across different institutions or learning environments. The relatively small sample size further reinforces this limitation. Consequently, future research should prioritise external validation using institution-specific LMS traces, temporally separated cohorts, and longitudinal learning data capable of capturing behavioural changes over time. Temporal validation is particularly important because student engagement behaviours may evolve dynamically across semesters, instructional modalities, and academic transitions.

Importantly, this study advances a broader understanding of responsible educational AI by demonstrating that predictive effectiveness, fairness, interpretability, and calibration reliability should be treated as complementary rather than competing objectives. The findings suggest that ethically grounded educational prediction systems are most valuable when they provide transparent and actionable insights capable of supporting human-centred educational decision-making. Rather than replacing educators, explainable machine learning systems should function as decision-support mechanisms that enhance institutional capacity to identify learning risks, personalise interventions, and promote equitable academic support structures. In this sense, the contribution of the study lies not only in demonstrating the predictive utility of behavioural engagement variables but also in advancing a deployment-oriented framework for trustworthy and fairness-aware AI in education.

CONCLUSION

This study demonstrates that ethically aligned and explainable machine learning models can effectively predict student academic performance using behavioural engagement and academic context variables without relying on sensitive demographic attributes. The findings reveal that engagement indicators, particularly visited_resources, raisedhands, and announcements_view, exerted substantially greater influence on prediction outcomes than structural academic characteristics, highlighting the pedagogical importance of sustained learner interaction within digital learning environments. Among the evaluated models, Random Forest achieved the strongest overall performance by combining high discriminative accuracy with superior calibration reliability, indicating that its probabilistic predictions aligned closely with observed student outcomes. The integration of SHAP and LIME further strengthened interpretability by revealing both global behavioural patterns and individual-level prediction mechanisms, thereby enhancing transparency and pedagogical trust in the predictive framework.

Importantly, this study advances a fairness-by-design approach to educational prediction by excluding demographic variables and prioritising actionable behavioural indicators capable of supporting ethically grounded intervention strategies. Rather than treating predictive performance and interpretability as competing objectives, the findings demonstrate that explainable ensemble models can simultaneously achieve technical robustness, fairness, and educational relevance. Essentially, this study contributes a practical and deployment-oriented framework for trustworthy AI-driven educational analytics capable of supporting proactive intervention, equitable student support, and evidence-based decision-making in digitally mediated learning environments.

Data Availability Statement

The dataset used in this research was obtained from Kaggle's open-access repository (https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data)

Conflict of Interest

Authors declare no conflict of interest.

Authorship Contribution Statement

Ayanwale: Conceptualization, design, data acquisition, data analysis/interpretation, critical revision of the manuscript

Chikezie: Drafting manuscript, editing/reviewing

REFERENCES

Adefemi, K. O., & Mutanga, M. B. (2025). A robust hybrid CNN–LSTM model for predicting student academic performance. *Digital*, *5*(2), 1-18. https://doi.org/10.3390/digital5020016 1-18

Ahmed, E. (2024). Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024(1), 1-15. . https://doi.org/10.1155/2024/4067721

Arévalo-Cordovilla, F. E. & Peña, M. (2025). Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. *Scientific Reports,* 15, 1-14. https://www.nature.com/articles/s41598-025-15388-9

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2023). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*, 55, 3026–3054. https://doi.org/10.3758/s13428-022-01939-9

Ayanwale, M. A., Kehinde, O., Mojoyinola, M. O., & Ezeh-Kehinde, C. (2026). Leveraging psychological data and transparent machine learning for student success prediction in digital learning systems. *Telematics and Informatics Reports*, 22, 100330. https://doi.org/10.1016/j.teler.2026.100330

Ayulani, I. D., Sarwinda, D., Ardaneswari, G., & Handari, B. D. (2023). Tree-based ensemble methods and their applications for predicting students' academic performance. *International Journal on Advanced Science, Engineering and Information Technology,* 13(3), 919. https://doi.org/10.18517/ijaseit.13.3.16880

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Baneres, D., Rodríguez, M. E., Guerrero, A. E., & Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences,* 10(13), 1-28. https://doi.org/10.3390/app10134427

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C.,Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N.,Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Educational Measurement: Issues and Practice, 42*(1), 58–68.https://doi.org/10.1111/emip.12595

Cabot, J. H. & Ross, E. G. (2023). Evaluating prediction model performance: discrimination, calibration and decision-analytic metrics. *Journal of Clinical Epidemiology.* 174(3), 723 726. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10529246/

Chen, T. Q. & Guestrin, C. (2016) Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

Deep Interpretable Ensembles. (2022). https://doi.org/10.48550/arxiv.2205.12729

Du Plooy, E., Casteleijn, D. & Franzsen, D. (2024). Personalized adaptive learning in higher education: A systematic review. *Heliyon, 10*(21), 1-24. https://doi.org/10.1186/s40561 02400350-5

Dunne, R., Reguant, R., Ramarao-Milne, P., Szul, P., Sng, L. M. F., Lundberg, M., Twine, N. A., & Bauer, D. C. (2023). Thresholding gini variable importance with a single trained random forest: An empirical bayes approach. *bioRxiv.* https://doi.org/10.1101/2022.04.06.487300

Fartuşnic, R. O., Istrate, O., &Fartusnic, C. (2025). Beyond automation: A conceptual framework for AI in educational assessment. *Journal of Digital Pedagogy,* 4(1), 83-102. https://digital-pedagogy.eu/conceptual-framework-for-ai-in-educational-assessment/

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1), 3. https://doi.org/10.3390/sci6010003

Garzón, J., Patiño, E., & Marulanda, C. (2025). Systematic review of artificial intelligence in education: Trends, benefits, and challenges. *Multimodal Technologies and Interaction*, *9*(8), 84. https://doi.org/10.3390/mti90800844

Gunasekara, S. & Saarela, M. (2025). Explainable AI in education: Techniques and qualitative assessment. *Applied Science*, 15, 1239. https://doi.org/10.3390/app15031239

Hoel, T., & Chen, W. (2018). Privacy and data protection in learning analytics should be motivated by an educational maxim—towards a proposal. Research and Practice in Technology Enhanced Learning, 13(1), 20. DOI: 10.1186/s41039-018-0086-8

Hooshyar, D. & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. *IEEE Access,* (20) 1-19. https://dio.org/10.1109/ACCESS.2024.3463948

Idowu, J. A. (2024). Debiasing education algorithms. Sahel AI. *International Journal of Artificial Intelligence in Education*, 34, 1510-1540. https://doi.org/10.1007/s40593-023-00389-4

Idowu, J. A., Koshiyama, A., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. Sahel AI. *Computes and Education: Artificial Intelligence*, 7, 1-13.https://doi.org/10.1016/j.caeai.2024.100267

Jiang, W. & Pardos, Z. A. (2021). Towards equity and algorithmic fairness in student grade prediction. *In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pp. 608–617, New York, college students. Association for Computing Machinery.

Kalita, E., Alfarwan, A. M., El Aouifi, H., Kukkar, A., Hussain, S., Ali, T., &Gaftandzhieva, S. (2025) Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. *Frontiers of Education,* 10, 1-18. http://doi.org/10.3389/feduc.2025.1581247

Kannan, R., Wen, C. H., Ng, Y. K. Ng, H., Yap, T. T. V., Wong, L. K., Chua, F. F., Goh, V. T., Lee, Y. L., & Wong, H. L. (2024). Predicting academic success and identifying at-risk students using ensemble and deep learning models. *Journal of System and Management Sciences,* 14(6), 93-112. https://doi.org/10.33168/JSMS.2024.0607

Kesgin, K., Kiraz, S., Kosunalp, S., &Stoycheva, B. (2025). Beyond performance: Explainingand ensuring fairness in student academic performance prediction with machine learning. *Applied Sciences,* 15, 1-25. https://doi.org/10.3390/app15158409

Kim, W. & Kim H. (2025). Counterfactual fairness evaluation of machine learning models on educational datasets. *arXiv,* 2, 1-21. https://doi.org/10.48550/arXiv.2504.11504

Kizilcec, R. F. and Lee, H. (2022) *Algorithmic Fairness in Education*. Routledge. https://doi.org/10.4324/9780429329067-10

Linardatos, P., Papastefanopoulos, V., &Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy* 2021, 23, 1-45. https://dx.doi.org/10.3390/e23010018

Liu, Q., Shakya, R., Khalil, M., & Jovanovic, J. (2025). Advancing privacy in learning analytics using differential privacy. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (LAK25). ACM. pp. 181-191. https://doi.org/10.1145/3706468.3706493

Lundberg, S.M. & Lee, S.-I. (2017) A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, 4-9 December 2017, 4766-4777.

Luo, J., Zheng, C., Yin, J. *et al.* (2025).Design and assessment of AI-based learning tools in higher education: A systematic review. *International Journal of Educational Technology in Higher Education,* 22, 1-27. https://doi.org/10.1186/s41239-025-00540-2

Lydersen, S. (2022). Logistic regression with more than two categories. Tidsskrift for Den Norske Laegeforening, 142, 10. https://doi.org/10.4045/tidsskr.21.0786

Matzavela, V., & Alepis, E. (2021). Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. 2, 100035. https://doi.org/10.1016/J.CAEAI.2021.100035

Movahedi, F., Padman, R., & Antaki, J. F. (2023). Limitations of receiver operating characteristiccurve on imbalanced data: Assist device mortality risk scores. *Journal of Thoracic Cardiovascular Surgery,* 165(4):1433-1442. https://doi.org/10.1016/j.jtcvs.2021.07.041.

Nunn, S. G., Avella, J. T., Kanai, T., &Kebritchi, M. (2016). Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. *Online Learning, 20*(2), 13-29. https://doi.org/10.24059/olj.v20i2.790

Organisation for Economic Co-operation and Development (OECD). (2023). *AI in education: Ensuring equity and inclusion*. OECD Publishing. https://doi.org/10.1787/ai-edu-2023

Rohman, M. G., Abdullah, Z., Kasim, S. & Rasyidah (2025). Hybrid logistic regression random forest on predicting student performance. *International Journal of Informatics. Visualization*, 9(2) 852-858. http://dx.doi.org/10.62527/joiv.9.2.3972

Svabensky, V., Verger, M., Rodrigo, M. M. T., Monterozo, C. J. G., Baker, R. S., Saavedra, S.,Lalle, M. Z. N. L., & Shimada, A. (2024). Evaluating algorithmic bias in models for predicting academic performance of filipino students. In B. PaaBen& C. D. Epp, (Editors), *Proceedings of the 17th International Conference on Educational Data Mining,* pp. 744–751.

Tang, Z., Jain, A., & Colina, F. E. (2024). A comparative study of machine learning techniques for college student success prediction. *Journal of Higher Education Theory and Practice*, 24(1), 101-116. https://doi.org/10.33423/jhetp.v24i1.6764

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. International Journal of Advanced Computer Science and Applications, 11(2), 612-619.

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, 15(3), 343. https://doi.org/10.3390/educsci15030343

Wang, D., Bian, C., & Chen, D. (2024). Using explainable AI to unravel classroom dialogue analysis. *British Journal of Educational Technology,* 55(6),2530-2556. https://bera journals.onlinelibrary.wiley.com/doi/full/10.1111/bjet.13466

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications,* 252, 124 167. https://doi.org/10.1016/j.eswa.2024.124167

Yacobson, E., Fuhrman, O., Hershkovitz, S., &Alexandron, G. (2021). De-identification is insufficient to protect student privacy, or What can a field trip reveal?. *Journal of Learning Analytics*, *8*(2), 83-92. https://doi.org/10.18608/jla.2021.7353

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology, 12,* 1-19. https://doi.org/10.3389/fpsyg.2021.698490

ISSN: 2790-4407