Babandi Usman, 2026, 14:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Explainable Student Failure Prediction: Convergent Multi-Method Predictor Identification and SHAP, LIME and ALE Interpretation of a Hybrid Ensemble

Babandi Usman[1], Abdullahi Hassan Birnin Kudu[2], Salim Ahmad[3], Zahraddeen Sufyan[4]
Department of Computer Science, Federal University Dutse[1]
Department of Cybersecurity, Federal University Dutse[2]
Department of Information Technology, Federal University Dutse[3]
Department of Computer Science, Federal University Dutse[4]

Abstract- Machine-learning models that predict student failure are often "black boxes": they flag a student as at risk without telling lecturers or advisers which factors drove the decision, which limits trust, validation and targeted intervention. Predictor relevance is also frequently established with a single feature-selection technique, and explainability is typically limited to one method (SHAP or LIME). This study identifies and ranks the most influential predictors of student academic failure using three complementary feature-selection techniques, and combines three post-hoc explainable-AI (XAI) methods SHAP, LIME and Accumulated Local Effects (ALE) to provide global, local and feature-effect explanations of a Hybrid Weighted Ensemble (Random Forest, XGBoost and ANN) for failure prediction. An educational dataset of 5,000 student records and 25 variables (2,807 non-failure and 2,193 failure cases) was prepared by median imputation (eight numerical variables, each <6% missing), categorical encoding and a stratified 80:20 split. Random Forest feature importance (embedded), Recursive Feature Elimination (wrapper) and Mutual Information (filter) were applied and their outputs compared. Random Forest, XGBoost, an ANN and their weighted probability ensemble (0.40/0.40/0.20; 0.50 threshold) were evaluated on 1,000 held-out students. SHAP (global and local attribution), LIME (local surrogate explanations) and ALE (global feature effects robust to correlated predictors) formed the explanation layer. Previous GPA was the most influential predictor in all analyses (Random Forest importance 0.114), followed by Attendance Rate, CA Score, Entry Score, Study Hours, Failed Courses, Assignment Score, LMS Logins and Materials Accessed. RFE retained ten predictors (these nine plus Age), and Mutual Information gave the strongest dependency with failure to Previous GPA, Failed Courses, Attendance Rate, Entry Score and Study Hours; five predictors were thus supported by all three methods. The models that were explained achieved 68.4% accuracy/75.4% ROC-AUC (Random Forest), 67.4%/74.1% (XGBoost), 63.1%/67.3% (ANN) and 68.7%/75.8% (Hybrid Weighted Ensemble). The SHAP summary ranked the predictors in the same order as the importance analysis; the LIME explanation decomposed an individual high-risk prediction into contributions from low Previous GPA, low attendance, failed courses, weak assessment and limited study effort; and ALE curves showed predicted failure risk falling with higher Previous GPA, attendance and study hours and rising with the number of failed courses. Failure risk in this dataset is associated with a combination of academic history, current assessment, attendance, study behaviour and digital engagement, and this finding is robust to the choice of feature-selection method. Using SHAP, LIME and ALE together answer three distinct questions which features matter, why a given student is flagged, and how predicted risk changes as a feature change and turns a prediction-only model into an interpretable decision-support tool. The explanations describe model behaviour and are not evidence of causation.

Keywords- explainable artificial intelligence; SHAP; LIME; accumulated local effects; feature selection; student failure prediction; educational data mining; ensemble learning

I. INTRODUCTION

Higher-education institutions now hold large volumes of student data from student information systems, learning management systems, assessment platforms and attendance records, and machine learning can use them to predict which students are likely to fail [1,2]. Highly accurate models ensembles,

© 2026 Babandi Usman. his is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

gradient-boosting machines and neural networks usually operate as black boxes whose internal reasoning is hard for educators, administrators and policymakers to follow. Although such a model may identify students at risk, it often cannot explain why they were classified as high risk or which factors contributed most, which reduces stakeholder confidence, limits adoption and hinders evidence-based intervention [3,4].

Explainable artificial intelligence (XAI) addresses this gap. Techniques such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME) and Accumulated Local Effects (ALE) identify the contribution of individual variables to model outputs at global and local levels [5,6]. In education, explanations let lecturers and advisers understand what influences student performance, justify and personalise interventions, detect potentially biased influences and strengthen accountability, so that predictions that cannot be explained are of limited practical value however accurate they are [3,7]. Two weaknesses remain in much educational XAI work. First, most studies use a single explainability technique typically SHAP or LIME and therefore supply either a global or a local view but rarely both, and explainability is often added as a supplementary analysis rather than being embedded in the prediction pipeline [8,9]. Second, the predictors that explanations are built on are frequently selected with one technique, so it is unclear whether they reflect the data or the selection method. Educational predictors are also typically correlated (attendance, assessment scores, study duration and online engagement), a situation in which some explanation methods are less reliable.

This paper addresses both weaknesses. Using a dataset of 5,000 students and 25 variables, it (i) identifies the influential predictors of failure with three complementary feature-selection techniques Random Forest importance, Recursive Feature Elimination (RFE) and Mutual Information (MI) and tests whether they converge; (ii) summarises the performance of the Random Forest, XGBoost, ANN and Hybrid Weighted Ensemble models whose behaviour is to be explained; and (iii) integrates SHAP, LIME and ALE into one explanation layer providing global, local and feature-effect views. The analysis is predictive and not causal: an influential predictor is not asserted to cause failure.

II. THEME

Explainability in educational prediction

XAI refers to computational methods that explain the reasoning of machine-learning algorithms in a form understandable to humans, so that users can understand, trust, validate and appropriately use AI systems. The aim is not to simplify the predictive model at the expense of accuracy but to explain complex models, which lets domain experts check whether predictions are logical, fair and consistent with existing knowledge [4,10]. Intrinsically interpretable models such as Logistic Regression or Decision Trees are transparent but often sacrifice accuracy on complex, nonlinear educational data; post-hoc methods instead analyse the behaviour of an already-trained model without altering it [6,11]. Because the framework studied here includes ensemble and neural models, post-hoc explanation is adopted.

SHAP, LIME and ALE

SHAP is derived from the Shapley value of cooperative game theory: each predictor is a "player" and its contribution is its average marginal contribution over all feature combinations, which gives mathematically consistent attributions [7]. SHAP yields global explanations (overall importance) and local explanations (contributions to an individual prediction). **LIME** explains one prediction by approximating the complex model around that observation with a simple interpretable surrogate, typically linear, and therefore concentrates on local interpretability. **ALE** is a global technique developed to overcome limitations of partial-dependence plots, which assume feature independence; ALE measures how incremental changes in a predictor influence the average prediction while accounting for correlations, which suits educational data in which attendance, assessment, study duration and online engagement are interrelated.

Because each method has distinct strengths and limitations, combining them gives a more complete

picture of model behaviour than any single one [3]: SHAP provides grounded feature importance, LIME explains individual cases and ALE shows overall feature effects. Related educational work has paired ensemble prediction with SHAP and LIME [8] or SHAP with stacked ensembles [9], and feature optimisation — removing redundant, irrelevant or noisy variables is known to improve accuracy, reduce overfitting and simplify interpretation [12,13].

III. MATERIALS AND METHODS

Data and preprocessing

The dataset comprises 5,000 student records and 25 variables: 14 numerical and 11 categorical variables, including identifier, demographic (Age, Gender, Location), admission (admission type, Entry Score), institutional (Institution, Department, Programme, Level, Semester, Academic Session, Region), academic-history (Previous GPA, Failed Courses, Credit Load), assessment (Assignment Score, CA Score), attendance (Attendance Rate), digital-engagement (LMS Logins, Materials Accessed, Quiz Attempts, Assignment Submissions) and study-behaviour (Study Hours) variables. The target, failure_status, is binary (0 = non-failure, 1 = failure) with 2,807 non-failure (56.14%) and 2,193 failure (43.86%) cases, and is separated from the predictors before training to prevent leakage. Missing values (below 6% in each of Age, Entry Score, Previous GPA, Assignment Score, CA Score, Attendance Rate, LMS Logins and Study Hours; 293–298 records each) were median-imputed; categorical variables were encoded numerically; and the data were split 80:20 with stratification into 4,000 training and 1,000 test records. Descriptive statistics showed marked variation in the variables of interest: mean Previous GPA 2.91 (0.64–5.00), mean Failed Courses 1.70 (0–7), mean Attendance Rate 73.50% (35–100%), mean Study Hours 8.36 (0–18.51) and mean LMS Logins 36.57 (5–69).

Feature selection

Three techniques with different perspectives were used (Table 1; Fig. 1). Random Forest importance is an embedded measure in which predictors producing greater impurity reduction receive higher importance. RFE is a wrapper that repeatedly evaluates importance and removes the least informative variables until the desired number is retained. Mutual Information is a filter that estimates the information shared between a predictor X and failure status Y, MI(X;Y); unlike linear correlation, it captures nonlinear dependence. Pearson correlation was used only as an exploratory diagnostic, because weak pairwise correlation does not imply low predictive value and nonlinear relations are missed, and it was not used to remove variables.

Table 1. Feature-selection methods used.

Method	Category	Main purpose	Strength
RF feature importance	Embedded	Estimate model-based importance	Captures nonlinear relationships
RFE	Wrapper	Identify an optimal predictive subset	Evaluates features in relation to model performance
Mutual Information	Filter	Measure predictor–target dependency	Captures nonlinear statistical relationships

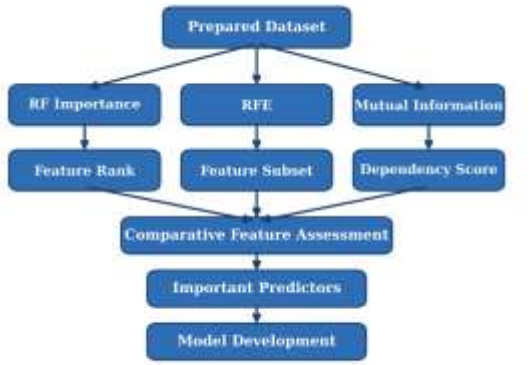


Fig. 1. Feature-selection flowchart.

Predictive models

Random Forest (bagging of decision trees), XGBoost (sequential gradient boosting) and a feed-forward Artificial Neural Network (input, hidden and output layers; $z = \Sigma(w_i x_i) + b$ with an activation function) were trained on the same training data and evaluated on the same held-out test set. Their failure probabilities were fused into the Hybrid Weighted Ensemble

$$P_{HWE} = 0.40\ P_{RF} + 0.40\ P_{XGB} + 0.20\ P_{ANN} \quad (1)$$

and a student was classified as failing when $P_HWE \geq 0.50$. Models were assessed with accuracy, precision, recall, F1-score, ROC-AUC and confusion matrices. The experiments were run in Python on Google Colaboratory with Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, SHAP, LIME and ALE-related tools.

Explainable AI layer

After prediction, the trained model's outputs are examined by SHAP, LIME and ALE (Fig. 2; Table 2). SHAP indicates whether a feature raises or lowers a student's predicted failure probability and by how much for example, a low Previous GPA or a high number of Failed Courses may increase predicted risk, while a high Attendance Rate may reduce it. LIME approximates the model around one student with an interpretable local surrogate and identifies the variables that most strongly support that student's predicted class. ALE shows how predicted failure probability changes across the range of an important predictor (Previous GPA, Attendance Rate, Study Hours, Entry Score, CA Score, Failed Courses). The three questions answered are: which features contribute most, and in what direction (SHAP); why was this particular student classified in this way (LIME); and how does the prediction change as an important feature change (ALE).

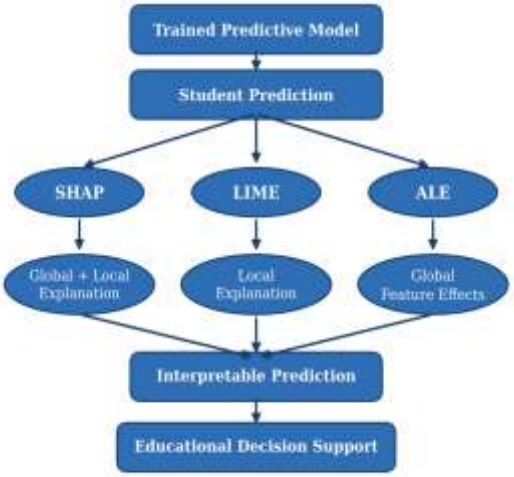


Fig. 2. Explainable-AI analysis flowchart.

Table 2. XAI techniques and their roles.

XAI technique	Explanation level	Principal function
SHAP	Global and local	Quantifies feature contribution
LIME	Local	Explains individual predictions
ALE	Global	Examines feature-effect relationships

IV. RESULTS AND DISCUSSION

Exploratory correlation analysis

The Pearson correlation matrix (Fig. 3) shows that previous-performance, assessment, attendance and engagement variables are associated with one another, as expected because academic behaviours and outcomes are interconnected; it also gives an initial indication of possible redundancy among predictors (for example, between Entry Score and

Previous GPA, and between LMS Logins and Materials Accessed). Because correlation alone cannot establish predictive value, it served only as a diagnostic step before the more robust feature-selection procedures.

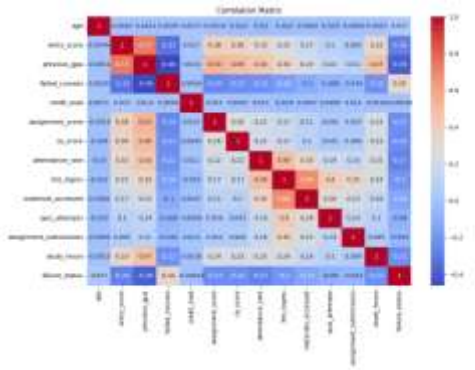


Fig. 3. Pearson correlation heatmap of the numerical study variables.

Random Forest feature importance

Random Forest importance identified Previous GPA as the most influential predictor, with an importance score of 0.114 (Fig. 4; Table 3). Attendance Rate, CA Score, Entry Score, Study Hours, Failed Courses and Assignment Score formed a second group of high-importance variables, followed by LMS Logins and Materials Accessed. The prominence of Previous GPA suggests that prior achievement is strongly associated with subsequent failure risk; Attendance Rate suggests that consistent participation matters; CA Score and Assignment Score indicate that continuous assessment carries information about ongoing progress; and LMS Logins and Materials Accessed show that digital engagement adds information beyond traditional academic records. Failure prediction should therefore not depend on examination outcomes alone.

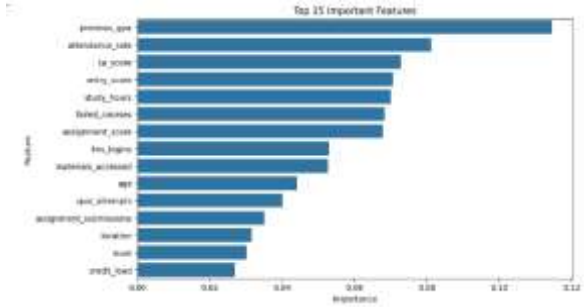


Fig. 4. Top 15 predictors by Random Forest feature importance.

Table 3. Principal predictors identified by Random Forest.

Rank	Predictor	Importance / relative influence
1	Previous GPA	0.114
2	Attendance Rate	High
3	CA Score	High
4	Entry Score	High
5	Study Hours	High
6	Failed Courses	High
7	Assignment Score	High
8	LMS Logins	Moderate–High
9	Materials Accessed	Moderate–High

Recursive Feature Elimination and Mutual Information

RFE selected ten predictors (Table 4): the nine variables above plus Age. The selected variables represent academic history (Previous GPA, Failed Courses), academic performance (CA Score, Assignment Score), admission background (Entry Score), engagement (Attendance Rate, Study Hours), digital learning behaviour (LMS Logins, Materials Accessed) and demographics (Age); a relatively compact set therefore captures much of the information needed for failure prediction. Mutual Information, which captures linear and nonlinear dependence, found the strongest dependency with failure status for Previous GPA, Failed Courses, Attendance Rate, Entry Score and Study Hours.

Table 4. Predictors selected by RFE.

Rank	Selected predictor	Rank	Selected predictor
1	Previous GPA	6	Failed Courses
2	Attendance Rate	7	Assignment Score
3	CA Score	8	LMS Logins
4	Entry Score	9	Materials Accessed
5	Study Hours	10	Age

Convergence across the three methods

Table 5 summarises which predictors each method supports. Previous GPA, Attendance Rate, Failed

Courses, Entry Score and Study Hours are supported by all three procedures; CA Score, Assignment Score, LMS Logins and Materials Accessed are supported by Random Forest and RFE; and Age only by RFE. This convergence gives stronger evidence that the identified predictors are not artefacts of a single selection technique and supports their use as the principal predictors in model development.

Table 5. Support for each predictor by the feature-selection methods.

Predictor	RF importance	RFE	MI (top five)
Previous GPA	✓	✓	✓
Attendance Rate	✓	✓	✓
Failed Courses	✓	✓	✓
Entry Score	✓	✓	✓
Study Hours	✓	✓	✓
CA Score	✓	✓	–
Assignment Score	✓	✓	–
LMS Logins	✓	✓	–
Materials Accessed	✓	✓	–
Age	–	✓	–

✓ = identified by the method; – = not reported among the method's principal predictors.

Performance of the models being explained

Explanations are only meaningful if the explained model has useful predictive ability. On the 1,000 test records, Random Forest achieved 68.4% accuracy and 75.4% ROC-AUC, XGBoost 67.4% and 74.1%, the ANN 63.1% and 67.3%, and the Hybrid Weighted Ensemble 68.7% accuracy, 66.9% precision, 56.7% recall, 61.4% F1-score and 75.8% ROC-AUC (Table 6). The ensemble's confusion matrix contained 438 true negatives, 123 false positives, 190 false negatives and 249 true positives. The ensemble's gain over Random Forest is modest (+0.3 pp accuracy, +0.4 pp ROC-AUC); the models discriminate well above chance (AUC > 0.5) but are far from perfect, and the recall of about 57% means many at-risk students are not flagged. This is why interpretation matters: the value of the framework lies in transparent, actionable information as well as in predictions.

Table 6. Test-set performance of the models.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	68.4 %	66.5 %	56.5 %	61.1 %	75.4 %
XGBoost	67.4 %	n/r	n/r	n/r	74.1 %
ANN	63.1 %	n/r	n/r	n/r	67.3 %
Hybrid Weighted Ensemble	68.7 %	66.9 %	56.7 %	61.4 %	75.8 %

SHAP: global and local attribution

SHAP provides feature-level attribution for the ensemble's predictions. The SHAP summary plot (Fig. 5) lists predictors by overall influence, with each point showing a student's SHAP value (impact on model output) coloured by the feature value. The ranking Previous GPA, Attendance Rate, Failed Courses, CA Score, Entry Score, Study Hours, Assignment Score, LMS Logins, Materials Accessed and Age is consistent with the feature-selection evidence, and Previous GPA is the strongest predictor. The plot also conveys the direction of each feature's contribution, its magnitude and its variation across students. Previous GPA should be discussed as the strongest predictor on the available evidence, not as a causal determinant of failure.

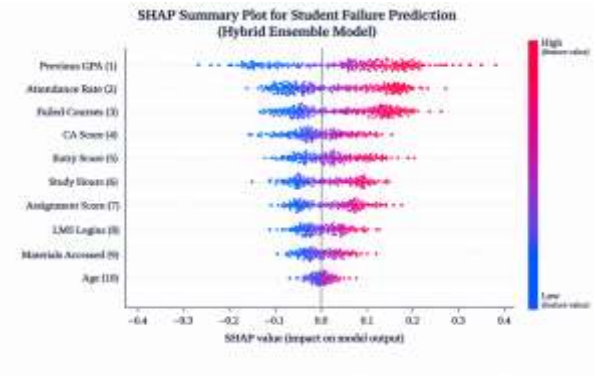


Fig. 5. SHAP summary plot for student failure prediction (Hybrid Weighted Ensemble).

LIME: explaining an individual prediction

LIME moves interpretation from population-level relationships to the individual student, which is what academic advisers need: not only which predictors matter in general, but why a particular student was classified as at risk. In the example in Fig. 6, a student predicted as high risk has the prediction supported mainly by a low Previous GPA, low attendance and failed courses, with additional weight from weak CA and assignment scores, low entry score and limited study hours; a small number of features act in the opposite direction. The explanation therefore points to a combination of academic-history, assessment and engagement factors, which suggests tailoring support (for example, engagement interventions for attendance-driven risk and tutoring or mentoring for persistent academic deficiency).

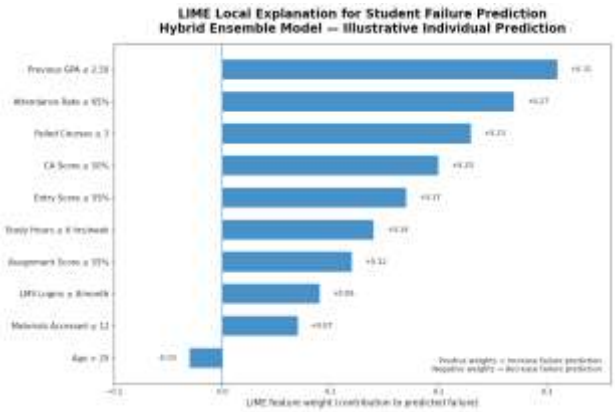


*Fig. 6. LIME local explanation of an individual prediction (illustrative individual prediction, Hybrid Weighted Ensem**ble).***

ALE: how predicted risk changes with predictor values

ALE shows how the model's predicted failure probability changes across the values of an explanatory variable, relative to the average prediction (Fig. 7). Predicted risk decreases as Previous GPA, Attendance Rate and Study Hours increase, and increases as the number of Failed Courses rises. These curves describe the behaviour of the predictive model with respect to the observed predictors and are not evidence of direct causality. Because ALE accounts for correlation among predictors, it is appropriate here, where attendance, assessment, study time and online engagement are interrelated.

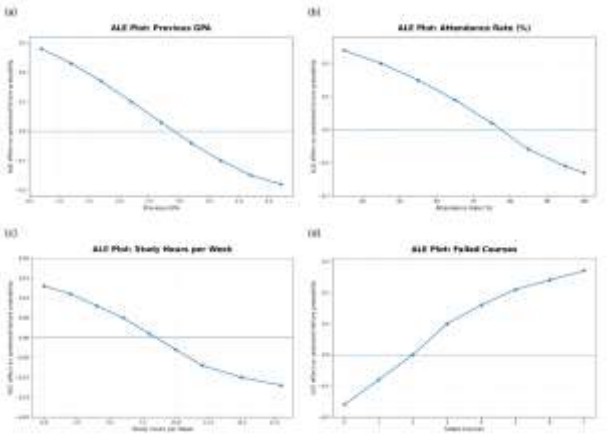


Fig. 7. ALE effects of key predictors on predicted failure probability: (a) Previous GPA, (b) Attendance Rate, (c) Study Hours and (d) Failed Courses.

Integrated interpretation and discussion

Taken together, the evidence gives a coherent picture (Table 7): failure is associated with a combination of academic history, current academic performance, engagement and learning behaviour, which reinforces the importance of multidimensional educational analytics.

Table 7. Integrated interpretation of the predictors.

Predictor	Interpretation in the predictive framework
Previous GPA	Strong indicator of prior academic performance
Attendance Rate	Represents participation and academic engagement
Failed Courses	Indicates previous academic difficulty
CA Score	Reflects continuous academic assessment
Entry Score	Represents admission-level academic preparation
Study Hours	Indicates learning effort
Assignment Score	Reflects coursework performance
LMS Logins	Indicates digital learning engagement
Materials Accessed	Reflects use of learning resources
Age	Provides demographic information

Previous academic performance. Previous GPA was the most consistently important predictor, indicating that historical performance contains substantial information about later outcomes; students with weaker records may need closer monitoring and additional support. Attendance and engagement. Attendance Rate, LMS Logins and Materials Accessed show that physical participation and digital engagement both carry information, so an effective early-warning system should use both. Failed courses. The number of previous failures was consistently important, which is intuitive because repeated difficulty may signal persistent learning challenges or insufficient mastery of prerequisites. Assessment performance. CA Score and Assignment Score were important, suggesting that continuous-assessment data can reveal difficulty before final examinations and thereby support earlier intervention.

The combination of SHAP, LIME and ALE is justified because no single technique answers all questions: SHAP shows which features matter and in what direction, LIME why a given student is flagged, and ALE how the model's prediction responds to a feature. Embedding the three in the prediction pipeline converts the framework from a prediction-only system into an explainable decision-support tool that answers "Why are they considered at risk?" in addition to "Which students are at risk?". Predictions should nevertheless be treated as decision-support information rather than automatic decisions, supporting lecturers, advisers and managers in investigating students' circumstances and choosing interventions.

V. CONCLUSION

Three complementary feature-selection techniques agreed that Previous GPA, Attendance Rate, Failed Courses, Entry Score and Study Hours are the most consistently influential predictors of student failure, with CA Score, Assignment Score, LMS Logins and Materials Accessed also important, and that failure reflects academic history, current assessment, attendance, study behaviour and digital engagement together. On this basis, a Hybrid Weighted Ensemble (68.7% accuracy; 75.8% ROC-AUC) was explained with SHAP, LIME and ALE, giving global, local and feature-effect views that can inform targeted, evidence-based academic support.

REFERENCES

1. Simon, P. D., Jiang, J., Fryer, L. K., King, R. B., & Frondozo, C. E. (2025). An assessment of learning management system use in higher education: Perspectives from a comprehensive sample of teachers and students. *Technology, Knowledge and Learning*, 30(2), 741–767. https://doi.org/10.1007/s10758-024-09734-5
2. Wang, Y. (2025). Artificial intelligence in student management systems to enhance academic performance monitoring and intervention. *Scientific Reports*, 15(1), 35122. https://doi.org/10.1038/s41598-025-19159-4
3. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., et al. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16. https://doi.org/10.1007/s12559-023-10179-8
4. Rane, J., Mallick, S., Kaya, Ö., & Rane, N. (2024). Enhancing black-box models: Advances in explainable artificial intelligence for ethical decision-making (pp. 136–180). https://doi.org/10.70593/978-81-981271-0-5_4
5. Wiggerthale, J., & Reich, C. (2024). Explainable machine learning in critical decision systems: Ensuring safe application and correctness. *AI*, 5, 2864–2896. https://doi.org/10.3390/ai5040138
6. Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. *Neural Processing Letters*, 57(1), 1–32. https://doi.org/10.1007/s11063-025-11732-2
7. Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications*, 36(9), 4847–4865. https://doi.org/10.1007/s00521-023-09232-2
8. Almazroei, E. E. (2026). Ensemble machine learning framework with SHAP and LIME for

accurate early prediction of student success in online learning environments. *Scientific Reports*.

9. Awadelkarim, A. M., & Albalawi, H. (2026). An explainable hybrid machine learning model for data-driven assessment and enhancement of program learning outcomes in higher education. *International Journal of Innovative Research and Scientific Studies*, 9(3), 84–105. https://doi.org/10.53894/ijirss.v9i3.11393
10. Obster, F., Ionita Ciolacu, M., & Humpe, A. (2024). Balancing predictive performance and interpretability in machine learning: A scoring system and an empirical study in traffic prediction. *IEEE Access*, 12, 195613–195626. https://doi.org/10.1109/ACCESS.2024.3521242
11. Mate, A. J. (2023). The role of explainability in user acceptance of artificial intelligence systems.
12. Kalita, E., Sunday, S., Gaftandzhieva, S., et al. (2025). Educational data mining: a 10-year review. *Discover Computing*. https://doi.org/10.1007/s10791-025-09589-z
13. 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, 698490. https://doi.org/10.3389/fpsyg.2021.698490