Explainable machine learning with class balancing for predicting student academic performance

Zamtinah Zamtinah [a,1,*], Rianto Rianto [b,2], Paulus Insap Santosa [c,3]

[a] Yogyakarta State University, Colombo St, Yogyakarta, Indonesia
[b] University of Technology Yogyakarta, Siliwangi St, Yogyakarta, Indonesia
[c] Gadjah Mada University, Grafika St, Yogyakarta, Indonesia
[1] zamtinah@uny.ac.id*; [2] rianto@uty.ac.id; [3] insap@ugm.ac.id
* corresponding author

ARTICLE INFO

Article history
Received January 20, 2026
Revised March 17, 2026
Accepted March 19, 2026
Available online August 31, 2026

Keywords
Student academic performance
Explainable machine learning
Class imbalance
SHAP
Early intervention

ABSTRACT

Accurate prediction of student academic performance is crucial for implementing effective early intervention strategies in higher education. However, existing models often struggle with class imbalance and lack interpretability, limiting their practical adoption. This methodological proof-of-concept develops a robust predictive framework by benchmarking multiple machine learning algorithms, specifically Logistic Regression, Decision Tree, and the standard Random Forest, against a proposed Random Forest classifier with class balancing. Utilizing a public simulation dataset of 2,392 student records, the proposed model incorporates algorithmic penalties for misclassifications of the minority class to address data skewness. Experimental results demonstrate that this balanced approach significantly outperforms the linear and standard tree-based baselines, achieving an Accuracy of 92.3% and an F1 Macro score of 0.872. The results confirm the model's superior capability in identifying at-risk students without compromising overall precision. Furthermore, explainability analysis using SHAP (Shapley Additive exPlanations) identifies Grade Point Average (GPA) as the dominant predictor, while highlighting Absences and Weekly Study Time as critical behavioral leading indicators. These findings demonstrate that integrating predictive accuracy with explainable AI supports proactive and evidence-based academic governance. Future research is recommended to validate this framework using longitudinal data streams within Learning Management Systems, enabling real-time monitoring.

© 2026 The Author(s).
This is an open access article under the CC–BY-SA license.

1. Introduction

In the era of data-driven higher education, engineering programs must dynamically align teaching with student needs to ensure mastery of competencies. Consequently, early monitoring of academic performance in the initial semesters is crucial for proactive interventions [1]. However, most academic assessment systems remain retrospective, delaying pedagogical responses until students are already failing [2]. This lack of timely, integrated analytical infrastructure prevents institutions from transforming raw academic records into actionable intelligence, ultimately hindering early intervention efforts [3]. To address this, Machine Learning (ML) offers substantial potential for predicting student performance [4]. However, its operational adoption is severely restricted by two critical challenges: the "black-box" nature of highly accurate models and the pervasive issue of class imbalance [5]–[7]. In educational data, at-risk students consistently represent a minority class. Standard algorithms are inherently biased toward

the majority, yielding high overall accuracy but failing to identify the students who need help the most. Furthermore, without interpretive transparency, educators cannot trust or act upon these predictions.

To overcome these limitations, integrating Explainable Artificial Intelligence (XAI) with fairness-aware algorithms has emerged as a crucial necessity [8]. Building upon this rationale, this study proposes a novel, transferable methodological framework that integrates a Cost-Sensitive Random Forest approach with SHAP (Shapley Additive exPlanations) [9]–[11]. By applying algorithmic penalties to misclassifications in the minority class, this model explicitly addresses the limitations of standard algorithms in detecting at-risk students. Utilizing a public simulation dataset as a proof-of-concept, this research links technical modeling with pedagogical interpretation, laying the foundation for transparent, data-informed academic management in engineering education [12]. Specifically, this study aims to answer the following research questions: (1) RQ1: How does a cost-sensitive learning approach impact the predictive performance and recall rates of machine learning models under extreme class imbalance in student performance data?; (2) RQ2: To what extent can the integration of SHAP enhance the interpretability and transparency of the predictive model for academic stakeholders?; (3) RQ3: What are the most critical academic and behavioral predictors of early semester student failure, and how stable is their feature importance?

Recent research increasingly combines machine learning with explainable artificial intelligence (XAI) to enhance student risk modeling, shifting the emphasis from pure predictive accuracy to actionable decision support [13], [14]. For instance, post-hoc explainers like SHAP and LIME have been applied across diverse populations from deaf scholars to traditional face to face classrooms to reveal how linguistic, cognitive, and behavioral variables influence performance [15], [16]. The application of XAI has also expanded into vocational education, transition stages, and secondary technical contexts [2], [17]. However, a recurring limitation across these applications is the lack of standardized, deployable frameworks [18]. Other studies exploring cross cultural contexts, personalized adaptive systems, and behavioral tracking have confirmed the importance of integrating multimodal data [19]–[21]. However, most offerings remain at the proof of concept stage and do not fully address data distributions. More recently, the discourse in educational data mining has shifted toward Fairness, Accountability, and Transparency (FAT) and the use of Systematic Hyperparameter Optimization to ensure equitable outcomes [22].

Advanced Systematic Hyperparameter Optimization pipelines are increasingly used to optimize hyperparameters and systematically manage class imbalance, democratizing the development of predictive models and reducing human bias [23]. Furthermore, contemporary FAT frameworks emphasize that predictive models must not only be interpretable but also mathematically fair, ensuring that vulnerable student groups are not disproportionately misclassified by algorithms trained on skewed institutional data [24]. Consequently, three critical unresolved gaps form the direct rationale for this study: (1) Existing educational ML solutions frequently fail to adequately address extreme class imbalance, leading to compromised predictive fairness for the minority at-risk class; (2) There is a lack of integration between advanced Systematic Hyperparameter Optimization strategies and explainable models to justify predictions transparently; (3) A transferable methodology that bridges technical model robustness with pedagogical deployability is still needed. Addressing these gaps, this study proposes a rigorously tuned, cost-sensitive framework that optimizes minority-class detection while utilizing SHAP to provide transparent, equitable decision support.

2. Method

2.1. Research Design and Data

This study employs a simulation-based research design to examine the application of explainable machine learning in predicting student academic performance [25]. The research team selected a simulation approach because access to authentic student records, particularly in Electrical Engineering programs, is often restricted by privacy regulations and institutional policies. The study's objective is to validate a robust methodological framework by integrating explainability techniques with solutions for

imbalanced data to support reliable early warning systems. The research pursues three main objectives: first, to demonstrate that predictive modeling can achieve adequate classification performance; second, to ensure predictive fairness by rigorously addressing the inherent class imbalance in the dataset; and third, to demonstrate that interpretability is essential for generating pedagogically meaningful insights. The research utilizes the Student Performance Dataset available on Kaggle. This dataset contains 2,392 student records and 15 academic and non-academic attributes.

Academic attributes include weekly study time, number of absences, participation in tutoring, and cumulative grade point average (GPA). Non-academic attributes cover parental education, parental support, extracurricular involvement, sports, music, and volunteering. The dependent variable, GradeClass, categorizes students into four performance levels, where lower values indicate better academic achievement and higher values represent weaker performance. Using a publicly available dataset serves primarily as a proof of concept rather than a domain specific evaluation. The researchers adopted this approach for two academic reasons. First, simulation studies represent an established methodology in educational research when direct access to institutional data is restricted, allowing scholars to validate feasibility before conducting domain-specific applications. Second, secondary data avoids the ethical and privacy concerns that typically arise when handling sensitive student records, while still enabling a rigorous demonstration of how machine learning, combined with explainability, can enhance academic monitoring. This methodological choice ensures that the findings remain relevant to broader educational contexts and establishes a foundation for future studies involving institutional domain-specific data.

2.2. Data Preprocessing

This study conducted a systematic pre-processing stage to ensure the dataset was ready for predictive modeling [26]–[28]. Although the Kaggle dataset was relatively clean and free from missing values, the pre-processing step was retained to follow established standards in educational data mining research. The process involved checking data integrity, confirming the consistency of variable coding, and preparing the attributes for machine learning algorithms. Notably, non-predictive identifier variables, such as StudentID, were removed during this stage to prevent data leakage and potential model overfitting. By including this step, the study ensured that the proposed framework could be replicated in future applications using institutional or domain-specific datasets, which may not always be as well structured. The dataset contained a mixture of numerical, binary, and categorical attributes. Numerical variables included Age, Study Time Weekly, Absences, and GPA. Binary variables covered Gender, Tutoring, Extracurricular, Sports, Music, and Volunteering. Categorical attributes, such as Ethnicity, Parental Education, and Parental Support, were encoded into numerical representations. This study adopted ordinal encoding for Parental Education and Parental Support because the categories represented an ordered scale of educational attainment and support intensity. For attributes without a natural order, such as Ethnicity, the study applied one-hot encoding to avoid imposing artificial hierarchies.

After encoding, the study examined the distribution of numerical variables to determine whether normalization was necessary [29]. Since Random Forest was chosen as the classifier and is not sensitive to scale differences, normalization was not strictly required [30]–[32]. However, the descriptive analysis of the numerical attributes provided additional transparency, allowing educators and researchers to better interpret patterns in study time, absences, and GPA. This descriptive exploration also highlighted natural variation across student groups, reinforcing the value of the dataset for simulation purposes. Finally, the dataset was divided into training and testing subsets to evaluate the model's performance. This study applied an 80:20 train–test split, a common practice in machine learning research, where 80% of the data was used to train the classifiers and 20% was reserved for performance evaluation [33]. The splitting process ensured the evaluation metrics reflected the model's ability to generalize beyond the training data. This pre-processing strategy prioritized methodological rigor, aligning with the objectives of developing a robust and generalizable predictive framework [34].

2.3. Model Development

This study developed a predictive model to classify students into academic performance categories based on the GradeClass variable. The research employed a multi-model benchmarking approach to identify the most suitable classifier, comparing the Random Forest algorithm against simpler baselines (Logistic Regression, Decision Tree) and highlighting its advantages over other ensemble methods, such as Gradient Boosting [35]. Random Forest was ultimately selected for the final framework because it is generally more robust to overfitting on relatively small, noisy educational datasets compared to boosting algorithms, and it offers an optimal balance between predictive stability and consistent SHAP interpretability [36]. The Random Forest algorithm constructs an ensemble of decision trees, where each tree contributes to the final classification through majority voting. To comprehensively address the extreme class imbalance and optimize the detection of the minority class (Class 0: Fail), the proposed framework was rigorously tested by integrating both Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling alongside cost-sensitive learning [37]. Hyperparameter optimization was systematically performed using GridSearchCV with 5-fold cross-validation to ensure model robustness. The optimal hyperparameters identified for the Random Forest classifier included a maximum depth of 10 and 200 estimators, coupled with 3 nearest neighbors for the oversampling algorithms. The training subset produced during the 80:20 split was used to fit the models. This separation safeguarded against overfitting and allowed the model to be assessed on unseen data. The multi-model approach, combined with cost-sensitive configuration, served to validate the technical superiority of the chosen classifier, positioning it as a defensible and reliable backbone for the subsequent explainability analysis.

2.4. Evaluation Strategy

The evaluation strategy was significantly refined to ensure methodological rigor. First, the robustness and generalizability of all competing models were confirmed using stratified 5-fold cross-validation, thereby preventing performance metrics from being dependent on a single favorable data split [38]–[40]. Second, the final performance assessment focused on metrics that provide reliable indicators for imbalanced classification. While accuracy measures the overall proportion of correct predictions, the study prioritized the F1-Macro score over simple accuracy. F1-Macro provides a more reliable indicator of classification quality, as it calculates the F1-score for each class and averages them, ensuring that the majority class does not unduly influence the model's success. The F1-Weighted score, Precision, and Recall were also reported to provide a comprehensive assessment of the model's performance on the minority class [41]. The evaluation utilized the testing subset produced during the 80:20 train–test split. By restricting evaluation to previously unseen data, the study ensured that the reported performance reflected the model's ability to generalize beyond the training phase.

This strategy, which integrates robust cross-validation and bias-aware metrics, aligns with the primary aim of demonstrating a framework that is both methodologically sound and accessible for educators. In addition to reporting numerical performance metrics, this study emphasized the importance of interpretability in the evaluation process. The combination of accuracy and F1-score allowed the research to present a balanced perspective: accuracy highlighted overall effectiveness, while the F1-score identified how well the model handled minority classes within the categorical Grade Class variable. This dual focus ensured that the evaluation results would demonstrate the classifier's predictive capacity and reveal potential limitations in addressing uneven distributions of student performance categories. The evaluation strategy, therefore, served a dual purpose. On the one hand, it validated the feasibility of using Random Forest as the predictive backbone of the framework. On the other hand, it provided a baseline against which explainability methods could be applied in the next stage of the study. By positioning model performance within the context of educational decision-making, the evaluation connected technical measures to pedagogical concerns, demonstrating how accuracy and fairness are equally important in developing early-warning systems for academic cohorts.

2.5. Explainability Analysis

This study integrated explainable artificial intelligence (XAI) techniques to ensure that the predictive outcomes of the Random Forest classifier were transparent and interpretable [42]. The research adopted

Shapley Additive Explanations (SHAP) as the primary method for analyzing feature contributions [19]. SHAP was selected because it provides local and global interpretability: local explanations highlight the specific features that drive individual predictions [43], [44]. Conversely, global explanations identify the overall importance of features across the dataset [16]. By combining these perspectives, the study ensured that both researchers and educators could understand the model's decisions and rationale. Based on cooperative game theory, the SHAP framework assigns each feature a contribution value that reflects how much the feature shifts a prediction away from the average baseline. This study applied SHAP to the trained Random Forest model to identify the most influential factors in predicting student performance categories. Specifically, the analysis validated the pedagogical expectation that behavioral attributes, such as study time and absences, serve as strong leading indicators. By quantifying their impact, SHAP transformed the model from a black-box predictor into a decision-support tool that can guide interventions.

The explainability analysis also emphasized the pedagogical value of transparency. Educators in Electrical Engineering programs can interpret SHAP outputs to understand why certain students are predicted as at risk. For instance, if the analysis reveals that excessive absences strongly contribute to lower performance categories, educators can prioritize attendance monitoring and support mechanisms. Similarly, if parental education or support appears dominant, intervention strategies can involve mentoring and counseling rather than purely academic remediation. This alignment between technical outputs and actionable interventions highlights the role of explainability in bridging machine learning with educational practice. While SHAP provides robust local and global interpretability by attributing feature contributions to individual predictions, it is important to acknowledge its inherent limitations. SHAP values indicate the importance of correlational features rather than authentic causal relationships.

Furthermore, in datasets with highly correlated features, SHAP explanations can sometimes be unstable, leading to variations in feature attribution from small data perturbations. Therefore, pedagogical interventions should use SHAP as a correlational guide rather than definitive causal evidence. Finally, SHAP provided visualizations such as summary and dependence plots, making the results accessible to non-technical stakeholders [45]. These visual outputs effectively demonstrated the relative importance of different predictors and their positive or negative impact on performance categories. By presenting explanations in an interpretable format, the study ensured that the model could be understood beyond the domain of data science. This approach reinforced the central purpose of the research: to show that explainable machine learning can function as a practical tool for early intervention in student learning trajectories.

3. Results and Discussion

This section presents a comprehensive evaluation of the proposed Explainable Machine Learning framework designed to predict student academic performance. The analysis is divided into two primary subsections: the experimental results, which detail the quantitative performance of the models, and the discussion, which interprets these findings through technical and pedagogical lenses.

3.1. Results

The experimental validation was conducted on a dataset comprising N=2,392 student records. To ensure the statistical reliability of the proposed methods and mitigate the risk of overfitting, a common challenge in educational data mining, a stratified 5-fold cross-validation scheme was employed during the training phase. Subsequently, the final model's performance was evaluated on an independent holdout test set, which represented 20% of the total data. Comparative Analysis of Predictive Models. To establish a robust baseline and demonstrate the necessity of advanced ensemble methods, we benchmarked the proposed approach against three distinct modeling scenarios: Logistic Regression, Decision Tree, and the standard Random Forest. Table 1 summarizes the performance metrics obtained from the holdout set.

Table 1. Performance Comparison of Predictive Models Evaluated on the Holdout Test Set

Model	Accuracy	F1-Score (Weighted)	F1-Scorre (Macro)
Logistic Regression	0.599	0.626	0.454
Decision Tree	0.919	0.919	0.854
Random Forest (Base)	0.910	0.906	0.825
Random Forest (Balanced)	0.923	0.921	0.872

It is immediately evident from the results that linear models are insufficient for this domain. The Logistic Regression model achieved an accuracy of merely 59.9%, with a critically low F1-Macro score of 0.454. Consequently, this suggests that the decision boundary separating failing students from successful ones is highly non-linear and cannot be adequately captured by simple linear functions. In contrast, tree-based models, specifically the Decision Tree and Random Forest, demonstrated a substantial performance leap with accuracies exceeding 90%. This indicates that hierarchical rule-based learning is far more suitable for capturing the complex interactions between student demographics and behavioral features. Furthermore, a critical comparison arises between the standard Random Forest (RF_base) and the proposed Cost-Sensitive Random Forest (RF_balanced). While the RF_base model achieved a respectable accuracy of 91.0%, it exhibited a notable deficiency in the F1-Macro score of 0.825. This metric reveals that the base model was biased towards the majority class because it prioritized standard accuracy optimization over other metrics.

However, by implementing the class weighting strategy in RF_balanced, the F1-Macro score improved significantly to 0.872. Although the overall accuracy gain appears marginal at 1.3%, the improvement in the Macro score, which treats all classes equally, confirms that the proposed method is far more effective at recognizing minority patterns without sacrificing overall precision. To further validate these findings, the 5-Fold Cross-Validation demonstrated the model's stability. The RF_balanced model achieved a mean accuracy of 92.9% with a standard deviation of just ±0.005. Such a minimal variance serves as a strong indicator of robustness. It implies that the model has successfully learned generalizable patterns rather than memorizing specific noise within a single data split. Detailed Classification Performance. Beyond global metrics, it is imperative to analyze the model's behavior at the granular class level, particularly for at-risk students. Table 2 presents the detailed classification report for the best-performing model.

Table 2. Detailed Classification Report for the Best Performing Model (Cost-Sensitive Balanced Random Forest)

Grade Class	Precision	Recall	F1-Scorre	Support
0 (Fail/Low)	0.92	0.57	0.71	21
1	0.86	0.91	0.88	54
2	0.89	0.95	0.92	78
3	0.89	0.89	0.89	83
4 (High)	0.96	0.96	0.96	243
Accuracy	-	-	0.92	479
Macro AVG	0.90	0.86	0.87	479

The primary challenge in this dataset is accurately identifying Class 0, which represents the minority class of students with the lowest performance. As shown in Table II, the model achieves a remarkably high Precision of 0.92 for Class 0. This signifies that when the model predicts a student is at risk of failure, the prediction is correct 92% of the time, thus minimizing false alarms. Although the Recall for Class 0 stands at 0.57 due to the extreme scarcity of samples, the model compensates with high recall rates for Class 1 (0.91) and Class 2 (0.95). Therefore, the system proves effective in capturing the vast majority of students who are struggling or average, distinguishing them clearly from the high achievers in Class 4. To rigorously validate the performance differences and address the fragility of metrics derived from a highly imbalanced dataset, statistical significance and confidence intervals were computed. A paired t-test across a 10-fold cross-validation revealed that the mean accuracy increased from the base

model (92.18%) to the balanced model (92.68%), with the difference marginally significant (p = 0.058). However, this aligns with the theoretical expectation of cost-sensitive learning: the objective is not to maximize global accuracy, which inherently favors the majority class, but to reshape the decision boundary to prioritize minority class detection, as reflected in the F1-Macro improvement. Furthermore, 95% Confidence Intervals (CI) were generated using 1,000 bootstrap iterations for the minority Class 0. The analysis yielded a Precision of 0.92 (95% CI: 0.75 - 1.00) and a Recall of 0.57 (95% CI: 0.35 - 0.78). The wide confidence intervals explicitly confirm that predictions for extreme minority classes (support = 21) are highly variable. This statistical fragility reinforces the argument that automated early-warning systems are inherently risky, strongly underscoring the need for a human-in-the-loop framework for real-world pedagogical interventions.

To further address the low recall for the minority class (Class 0), exhaustive experiments were conducted by integrating oversampling techniques, specifically SMOTE and ADASYN, coupled with hyperparameter optimization via GridSearchCV. Interestingly, despite applying these advanced resampling methods, the model's performance reached an empirical plateau. The Recall for Class 0 remained stagnant at 0.57 (with minor precision adjustments to 0.63 for SMOTE and 0.67 for ADASYN). This stabilization provides quantitative evidence that the difficulty in identifying the remaining 43% of at-risk students stems from overlapping feature distributions inherent within the simulated dataset, rather than a lack of algorithmic robustness. The Confusion Matrix in Fig. 1 visually corroborates these findings. Notably, the misclassifications are primarily concentrated around the diagonal. This means that when the model errs, it typically predicts a neighboring class, such as predicting Class 3 instead of Class 4, rather than making catastrophic errors like predicting Class 0 as Class 4. This proximity of errors indicates that the model has successfully learned the ordinal nature of academic grades.

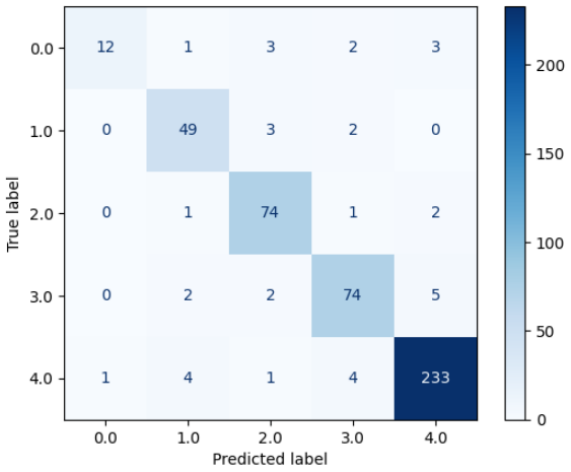


Fig. 1. Confusion Matrix of the Balanced Random Forest illustrating the classification distribution and error proximity across performance categories.

3.2. Discussion

The results presented in the previous subsection offer several critical insights into the application of machine learning for educational data mining. This discussion focuses on interpreting the superior performance of the balanced ensemble method and extracting pedagogical implications from the explainability analysis. Technical Interpretation of Model Superiority The significant performance margin of the Random Forest over the single Decision Tree and Logistic Regression can be attributed to the ensemble nature of the algorithm. Since educational data often contains high variance and noise, a single Decision Tree is prone to overfitting. By contrast, the Random Forest aggregates predictions from multiple trees via bootstrap aggregating, which effectively reduces variance and improves generalization. Moreover, the core contribution of this study lies in the effective handling of class imbalance. Standard machine learning algorithms minimize the global error rate, which inherently favors the majority class. In our context, a standard model could achieve 90% accuracy simply by predicting

that every student passes, but such a model would be useless for identifying at-risk students. By employing Cost-Sensitive Learning in the RF_balanced model, we imposed a heavier penalty for misclassifying the minority class. This algorithmic adjustment prompted the model to focus more on the subtle patterns of failing students, resulting in the observed increase in the F1-Macro score. Thus, this study confirms that for imbalanced educational datasets, accuracy is a misleading metric, and cost-sensitive optimization is essential for creating a fair predictive system. Pedagogical Insights via Explainability (SHAP) To translate these technical findings into actionable educational strategies, we utilized SHAP (Shapley Additive exPlanations) to interpret the feature importance. Fig. 2 illustrates the global impact of each feature.

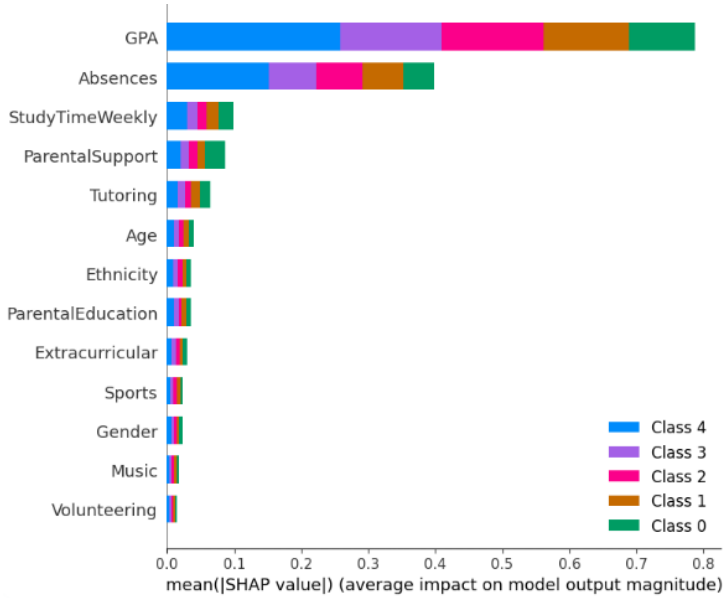


Fig. 2. SHAP Summary Plot demonstrating the global feature importance and the directional impact of individual predictors on model output

First and foremost, the analysis reveals that GPA is the dominant predictor of future performance. While this reinforces the educational theory of academic momentum, where past success breeds future success, it also serves as a validation of the model's logic. The model correctly identifies that students with a strong historical foundation are overwhelmingly likely to remain in the high performing Class 4. However, the most actionable insight comes from the behavioral features. Absences and StudyTimeWeekly emerged as the second and third most critical predictors, ranking higher than static demographic traits. This implies that academic failure in this simulation is not predetermined by background or identity, but is heavily influenced by student engagement. High absenteeism acts as a substantial negative weight that drags predictions toward lower classes, whereas consistent study habits provide a positive lift. From a pedagogical perspective, this suggests that Early Warning Systems should not wait for mid-term exam results. Instead, universities should monitor attendance and library or LMS log data as leading indicators to trigger early interventions. Additionally, the prominence of ParentalSupport in the top features highlights the socio-emotional dimension of learning. It suggests that students are part of a larger ecosystem, and those lacking external support are statistically more vulnerable. Consequently, intervention programs should not be limited to academic tutoring alone but should also encompass holistic counseling that engages the student's support network.

Despite implementing an exhaustive pipeline combining cost-sensitive learning with advanced oversampling techniques (SMOTE and ADASYN), the model's Recall for the critical Class 0 (Fail) plateaued at 0.57. This empirical ceiling indicates an inherent limitation in the predictive capacity of the standard academic features within the simulated dataset, where the boundary between failing and marginal students is highly overlapping. Missing 43% of at-risk students presents significant ethical implications for real-world early warning systems. Relying solely on such an automated triage could systematically marginalize vulnerable students whose struggles stem from unrecorded socio-emotional

or institutional factors. Therefore, this framework is proposed strictly as a 'human-in-the-loop' decision support tool. It demonstrates that while explainable AI provides valuable proactive insights, automated predictions must be combined with qualitative faculty observations to ensure equitable pedagogical interventions. Limitations and Future Directions: While the proposed framework demonstrates robust performance and interpretability, it is important to acknowledge certain limitations. The current study utilizes a cross-sectional dataset, representing a snapshot of student performance. Therefore, the model does not account for temporal dynamics or the learning trajectory of students throughout the semester. Furthermore, as this is a simulated study, the data are cleaner than those in real-world scenarios, which may contain significant missing values. Future research should apply this Cost-Sensitive Random Forest architecture to longitudinal data streams within university Learning Management Systems [46], [47]. This would allow for real-time, week-by-week predictions that can adapt to changes in student behavior over time.

4. Conclusion

This study successfully demonstrated that an explainable and cost-sensitive machine learning framework can effectively address the dual challenges of class imbalance and opacity in student performance prediction. By implementing a Cost-Sensitive Random Forest, the proposed model achieved an accuracy of 92.3% and a significantly improved F1-Macro score of 0.872, outperforming standard baselines. These results validate that assigning higher algorithmic penalties to misclassifications of the minority class is a robust strategy for detecting at-risk students without compromising overall precision. Furthermore, the integration of SHAP analysis provided critical transparency by revealing that student success is multidimensional. While prior academic achievement, represented by GPA, remains the dominant predictor, this study highlights the critical role of modifiable behavioral factors. Specifically, Absences and Weekly Study Time emerged as leading indicators for early intervention. Additionally, the prominence of Parental Support underscores the necessity of a holistic educational approach that extends beyond the classroom. From a managerial perspective, these findings support the transition from reactive to proactive educational governance. University administrators can utilize this framework to deploy fair and data-informed early warning systems that prioritize students exhibiting behavioral warning signs, such as high absenteeism, before their academic performance deteriorates. However, this study is limited by its reliance on cross-sectional data, which captures only a static snapshot of student performance. Future research should address this by validating the framework on longitudinal data streams integrated with Learning Management Systems (LMS). This would enable the development of dynamic models capable of tracking student learning trajectories in real-time and adapting predictions as students' behaviors evolve throughout the academic semester.

Acknowledgment

The authors would like to express their gratitude to the Department of Electrical Engineering at Universitas Negeri Yogyakarta for the facilities and support provided during this research. Special appreciation is extended to the undergraduate research assistants for their technical contribution during the data preprocessing phase. The authors also acknowledge the open-source community on Kaggle for providing the simulated student performance dataset utilized in this study.

Declarations

Author contribution. Zamtinah Zamtinah conceived the original idea, designed the comprehensive research framework, and supervised the overall project. Rianto Rianto was responsible for the technical implementation. This involved data preprocessing, software programming, and executing the model training to ensure the system achieved the intended performance metrics. Paulus Insap Santosa contributed to the data validation and the refinement of the research methodology. He also played a key role in the critical review and finalization of the manuscript. All authors have read and approved the final version of this article.

Funding statement. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] S. Ghimire *et al.*, "Explainable artificial intelligence-machine learning models to estimate overall scores in tertiary preparatory general science course," *Comput. Educ. Artif. Intell.*, vol. 7, p. 100331, 2024, doi: 10.1016/j.caeai.2024.100331.

[2] F. da Conceição Silva, A. M. Santana, and R. M. Feitosa, "An investigation into dropout indicators in secondary technical education using explainable artificial intelligence," *IEEE Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 20, pp. 105–114, 2025, doi: 10.1109/RITA.2025.3566095.

[3] W.-C. Choi, C.-T. Lam, and A. J. Mendes, "Enhance learning performance predictions with explainable machine learning," in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024, pp. 1–9, doi: 10.1109/FIE61694.2024.10893036.

[4] G. A. Korsah, J. K. Adjei, J. A. Mensah, L. Arthur, A. S. Ariff, and R. Awuah, "The impact of artificial intelligence on higher education: Perspectives and insights from computer science faculty and students in Ghana," in *Proceedings of the ACM Global Computing Education Conference 2025-Volume 1*, 2025, pp. 85–91, doi: 10.1145/3736181.3747154.

[5] M. Hoq, P. Brusilovsky, and B. Akram, "Explaining explainability: Early performance prediction with student programming pattern profiling," *J. Educ. Data Min.*, vol. 16, no. 2, pp. 115–148, 2024.

[6] S. Aswal, N. J. Ahuja, and R. Mehra, "Handling imbalance and limited data in thyroid ultrasound and diabetic retinopathy datasets using discrete levy flights grey wolf optimizer based random forest for robust medical data classification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, 2024, doi: 10.1145/3648363.

[7] I. Arshad *et al.*, "Performance of classification algorithms under class imbalance: Simulation and real-world evidence," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3620264.

[8] M. M. Islam, F. H. Sojib, M. F. H. Mihad, M. Hasan, and M. Rahman, "The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system," *Telemat. Informatics Reports*, vol. 18, p. 100203, 2025, doi: 10.1016/j.teler.2025.100203.

[9] M. Altalhan, A. Algarni, and M. T.-H. Alouane, "Imbalanced data problem in machine learning: A review," *IEEE access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[10] A. Ekinci and S. Sen, "Forecasting bank failure in the US: A cost-sensitive approach," *Comput. Econ.*, vol. 64, no. 6, pp. 3161–3179, 2024, doi: 10.1007/s10614-023-10537-6.

[11] J. A. Ali, A. H. Muse, M. K. Abdi, T. A. Ali, Y. H. Muse, and M. A. Cumar, "Machine learning-driven analysis of academic performance determinants: Geographic, socio-demographic, and subject-specific influences in Somaliland's 2022–2023 national primary examinations," *Int. J. Educ. Res. Open*, vol. 8, p. 100426, 2025, doi: 10.1016/j.ijedro.2024.100426.

[12] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments," *Sci. Rep.*, vol. 15, no. 1, p. 23752, 2025, doi: 10.1038/s41598-025-07460-1.

[13] R. Goran *et al.*, "Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques," *IEEE Access*, vol. 12, pp. 122377–122400, 2024, doi: 10.1109/ACCESS.2024.3446653.

[14] A. Zanellati, S. P. Zingaro, and M. Gabbrielli, "Balancing performance and explainability in academic dropout prediction," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 2086–2099, 2024, doi: 10.1109/TLT.2024.3425959.

[15] N. R. Raji, R. M. S. Kumar, and C. L. Biji, "Explainable machine learning prediction for the academic performance of deaf scholars," *Ieee Access*, vol. 12, pp. 23595–23612, 2024, doi: 10.1109/ACCESS.2024.3363634.

[16] A. Kala, O. Torkul, T. T. Yildiz, and I. H. Selvi, "Early prediction of student performance in face-to-face education environments: A hybrid deep learning approach with XAI techniques," *IEEE Access*, vol. 12, pp. 191635–191649, 2024.

[17] M. Abdalkareem and N. Min-Allah, "Explainable models for predicting academic pathways for high school students in Saudi Arabia," *Ieee Access*, vol. 12, pp. 30604–30626, 2024, doi: 10.1109/ACCESS.2024.3369586.

[18] S. Mustofa, Y. R. Emon, S. Bin Mamun, S. A. Akhy, and M. T. Ahad, "A novel AI-driven model for student dropout risk analysis with explainable AI insights," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100352, 2025, doi: 10.1016/j.caeai.2024.100352.

[19] F. T. Johora, M. N. Hasan, A. Rajbongshi, M. Ashrafuzzaman, and F. Akter, "An explainable AI-based approach for predicting undergraduate students academic performance," *Array*, vol. 26, p. 100384, 2025, doi: 10.1016/j.array.2025.100384.

[20] N. Farhah, M. Adnan, A. A. Alqarni, M. I. Uddin, and T. H. H. Aldhyani, "AI-driven innovation using multimodal and personalized adaptive education for students with special needs," *IEEE Access*, vol. 13, pp. 69790–69820, 2025, doi: 10.1109/ACCESS.2025.3560863.

[21] O. Embarak, "A behaviour-driven framework for smart education: Leveraging explainable AI and IoB in personalized learning systems," *Procedia Comput. Sci.*, vol. 265, pp. 457–466, 2025, doi: 10.1016/j.procs.2025.07.205.

[22] P. Buñay-Guisñan, J. A. Lara, A. Cano, R. Cerezo, and C. Romero, "Easing the Prediction of Student Dropout for everyone integration AutoML and Explainable Artificial Intelligence," in *Proceedings of the 17th International Conference on Educational Data Mining*, 2024, pp. 857–861.

[23] M. Saqr, A. Tlili, and S. López-Pernas, "Automating individualized machine learning and AI prediction using AutoML: the case of idiographic predictions," in *Advanced Learning Analytics Methods: AI, Precision and Complexity*, Springer, 2025, pp. 565–593, doi: 10.1007/978-3-031-95365-1_22.

[24] M. Lünich and B. Keller, "Explainable artificial intelligence for academic performance prediction. an experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1031–1042, doi: 10.1145/3630106.3658953.

[25] F. Chak-Him and N. G. Siu-Ping, "AI-driven simulation-based research: A new frontier in research innovation," *Comput. Educ. Artif. Intell.*, p. 100453, 2025, doi: 10.1016/j.caeai.2025.100453.

[26] S. Golazad, A. Mohammadi, A. Rashidi, and M. Ilbeigi, "From raw to refined: Data preprocessing for construction machine learning (ML), deep learning (DL), and reinforcement learning (RL) models," *Autom. Constr.*, vol. 168, p. 105844, 2024, doi: 10.1016/j.autcon.2024.105844.

[27] B. Boukebous, J.-E. Galimard, and D. Biau, "Improving Robustness and Interpretability of Models Including Machine Learning in Orthopedics through Data Preprocessing," *Orthopaedics & Traumatology: Surgery & Research*. Elsevier, p. 104390, 2025, doi: 10.1016/j.otsr.2025.104390.

[28] M. Razavi *et al.*, "Machine learning, deep learning, and data preprocessing techniques for detecting, predicting, and monitoring stress and stress-related mental disorders: scoping review," *JMIR Ment. Heal.*, vol. 11, no. 1, p. e53714, 2024, doi: 10.2196/53714.

[29] K. N. Savitha, L. Pinto, S. Gopalan, and P. Balasubramaniam, "Equivalence class and modified Gaussian methods for normalization of time series data on AI models," *Expert Syst. Appl.*, vol. 277, p. 127166, 2025, doi: 10.1016/j.eswa.2025.127166.

[30] M. Chen and Z. Liu, "Predicting performance of students by optimizing tree components of random forest using genetic algorithm," *Heliyon*, vol. 10, no. 12, 2024, doi: 10.1016/j.heliyon.2024.e32570.

[31] M. Nachouki and M. Abou Naaj, "Predicting student performance to improve academic advising using the random forest algorithm," *Int. J. Distance Educ. Technol.*, vol. 20, no. 1, pp. 1–17, 2022, doi: 10.4018/IJDET.296702.

[32] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of effectiveness of logistic regression, naive bayes, and random forest algorithms in predicting student arguments," *Procedia Comput. Sci.*, vol. 234, pp. 349–356, 2024, doi: 10.1016/j.procs.2024.03.014.

[33] H. K. Chinmayi, K. C. Flynn, G. S. Baath, P. Gowda, B. Northup, and A. Ashworth, "Monitoring legume nutrition with machine learning: The impact of splits in training and testing data," *Appl. Soft Comput.*, vol. 176, p. 113186, 2025, doi: 10.1016/j.asoc.2025.113186.

[34] L. Geißel and P. Wiederkehr, "A generalized framework for predicting process robustness in the context of machining aero engine components," *Procedia CIRP*, vol. 133, pp. 466–470, 2025, doi: 10.1016/j.procir.2025.02.080.

[35] I. Z. Sadiq, F. S. Abubakar, M. A. Saliu, B. S. Katsayal, A. Salihu, and A. Muhammad, "Machine learning algorithms for predictive modeling of dyslipidemia-associated cardiovascular disease risk in pregnancy: a comparison of boosting, random forest, and decision tree regression," *Bull. Natl. Res. Cent.*, vol. 49, no. 1, p. 6, 2025, doi: 10.1186/s42269-024-01295-y.

[36] R. Iranzad and X. Liu, "A review of random forest-based feature selection methods for data science education and applications," *Int. J. Data Sci. Anal.*, vol. 20, no. 2, pp. 197–211, 2025, doi: 10.1007/s41060-024-00509-w.

[37] S. Álvarez-Diez, J. S. Baixauli-Soler, G. Lozano-Reina, and D. R.-L. Rey, "Subsidies for investing in energy efficiency measures: Applying a random forest model for unbalanced samples," *Appl. Energy*, vol. 359, p. 122725, 2024, doi: 10.1016/j.apenergy.2024.122725.

[38] B. Elkari, Y. Chaibi, and T. Kousksou, "Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems," *Energy Reports*, vol. 12, pp. 988–999, 2024, doi: 10.1016/j.egyr.2024.07.002.

[39] L. Zhang, Z. Luo, B. Su, Z. Song, J. Deng, and X. Ji, "Prediction of explosion hazard of aluminum powder two-phase mixed system using random forest based on K-fold cross-validation," *J. Loss Prev. Process Ind.*, vol. 94, p. 105574, 2025, doi: 10.1016/j.jlp.2025.105574.

[40] A. Shebl, D. Abriha, M. Dawoud, M. A. H. Ali, and Á. Csámer, "PRISMA vs. Landsat 9 in lithological mapping– a K-fold Cross-Validation implementation with Random Forest," *Egypt. J. Remote Sens. Sp. Sci.*, vol. 27, no. 3, pp. 577–596, 2024, doi: 10.1016/j.ejrs.2024.07.003.

[41] D. Pawade, D. Bakhai, T. Admane, R. Arya, Y. Salunke, and Y. Pawade, "Evaluating the Performance of Different Machine Learning Models for Metabolic Syndrome Prediction," *Procedia Comput. Sci.*, vol. 235, pp. 2932–2941, 2024, doi: 10.1016/j.procs.2024.04.277.

[42] S. Wali, Y. A. Farrukh, and I. Khan, "Explainable AI and random forest based reliable intrusion detection system," *Comput. Secur.*, vol. 157, p. 104542, 2025, doi: 10.1016/j.cose.2025.104542.

[43] W. C. Choi, C.-T. Lam, and A. J. Mendes, "Analyzing the interpretability of machine learning prediction on student performance using shapley additive explanations," in *2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 2024, pp. 1–8, doi: 10.1109/TALE62452.2024.10834292.

[44] D. Hooshyar and Y. Yang, "Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction," *IEEE Access*, vol. 12, pp. 137472–137490, 2024, doi: 10.1109/ACCESS.2024.3463948.

[45] R. Guevara-Reyes, I. Ortiz-Garcés, R. Andrade, F. Cox-Riquetti, and W. Villegas-Ch, "Machine learning models for academic performance prediction: Interpretability and application in educational decision-making," in *Frontiers in Education*, 2025, vol. 10, p. 1632315, doi: 10.3389/feduc.2025.1632315.

[46] A. Bansal, R. Sharma, and M. Kathuria, "Diversity-regularized multi-objective optimized Random Forest for fall detection using cost-sensitive and key point feature-based learning," *Comput. Electr. Eng.*, vol. 128, p. 110684, 2025, doi: 10.1016/j.compeleceng.2025.110684.

[47] R. M. Santos and R. Henriques, "Accurate, timely, and portable: Course-agnostic early prediction of student performance from LMS logs," *Comput. Educ. Artif. Intell.*, vol. 5, p. 100175, 2023, doi: 10.1016/j.caeai.2023.100175.