iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81669

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Explainable Student Performance Prediction using XGBoost: A Stability-Based Comparative Study of SHAP and LIME

Kashish Tomar, Anshu Jain, Dr. Rinky Ahuja
School of Engineering and Technology Sushant University Gurugram, India

Abstract: ***Predicting student academic success is essential for providing early support to at-risk learners. However, high-accuracy models such as Extreme Gradient Boosting (XGBoost) are often underutilized by educators due to their opaque decision-making processes. This paper implements a predictive framework using the UCI Student Performance dataset (649 records). We evaluate Random Forest, XGBoost, Logistic Regression, and a Multi-Layer Perceptron (MLP) baseline, with our XGBoost model achieving a classification accuracy of 0.892 and an F1-score of 0.936. To provide transparency, we integrate SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) as Explainable AI (XAI) methods. Our analysis shows that previous academic results (G2) and past failures are the most significant predictors. We quantitatively compare these XAI methods, finding that SHAP demonstrates near-perfect stability (≈1.000) compared to LIME (0.846). A demographic parity evaluation confirms equitable prediction across socioeconomic groups. This study provides a practical framework for educational decision support, integrating predictive performance with the interpretability required for institutional trust***

Keywords: Educational data mining, explainable AI, imbalanced classification, machine learning, SHAP, student performance prediction, XGBoost

I. INTRODUCTION

Educational Data Mining (EDM) has evolved from descriptive reporting tools toward the construction of predictive early-warning systems [9]. Identifying students at risk of academic failure has become feasible through the analysis of demographic, social, and behavioral data. However, as predictive models grow in complexity, their opaque, "black-box" nature introduces a trust barrier. Educators are frequently reluctant to act on model predictions when the underlying reasoning cannot be adequately explained [4].

Our work addresses this by developing a machine learning pipeline that predicts student outcomes (Pass/Fail) while explaining the logic behind those predictions. We focus on the UCI Student Performance dataset [1] and utilize SHAP [5] and LIME [6] to extract both global and local insights. We also evaluate the model for demographic parity to ensure predictions do not unfairly penalize students based on socioeconomic factors such as parental education level.

The primary contribution of this study is a quantitative stability-based evaluation of SHAP and LIME within an educational risk-prediction framework, providing empirical evidence on explanation consistency—an aspect often overlooked in prior EDM research. While prior studies have applied machine learning to student performance prediction [2, 3], limited work has quantitatively evaluated the stability of explainability methods within educational settings. This study addresses this gap by introducing a stability-based comparison of SHAP and LIME.

This study pursues four primary research objectives:

1) Benchmarking: Compare the performance of XGBoost, Random Forest, Logistic Regression, and an MLP neural baseline on the UCI Portuguese student dataset.
2) Explainability: Apply SHAP for global feature importance and LIME for local, instance-level diagnostics.
3) Stability Testing: Quantify the reliability of XAI explanations by measuring consistency across multiple runs using Spearman rank correlation.
4) Feature Impact: Differentiate between mutable features (e.g., absences) and immutable features (e.g., mother's education) to suggest actionable interventions.

©IJRASET: All Rights are Reserved

The remainder of this paper is organized as follows: Section II reviews related work on EDM and XAI methods. Section III details the dataset, preprocessing, and implementation. Section IV presents experimental results and explainability analysis. Section V discusses key findings and their implications. Section VI outlines limitations, and Section VII concludes with directions for future work.

II. RELATED WORK

A. Dataset Standards

The foundations of EDM were established by Romero and Ventura [9], who highlighted the need for tools that support pedagogical decision-making. The UCI Student Performance dataset, introduced by Cortez and Silva [1], remains a primary benchmark. Their research established that while early-term grades (G1, G2) are strong predictors of the final grade (G3), using them can make the prediction task trivial. Recent work by Durães et al. [16] compared five classifiers on the same UCI dataset, finding SVM competitive but noting the critical role of feature selection—reinforcing the need for XAI-based feature attribution. Similarly, Frontera et al. [17] demonstrated that XGBoost with SHAP achieves strong generalizability across geographically diverse student populations, supporting the cross-domain applicability of our approach.

Kuzilek et al. [10] introduced the Open University Learning Analytics Dataset (OULAD), a large-scale alternative benchmark. However, the UCI dataset remains widely used due to its tractability and rich set of demographic and behavioral features. Breiman's Random Forest [8] and Chen and Guestrin's XGBoost [7] are the two dominant ensemble learners applied to this domain.

B. Explainable AI Trends

Recent research has shifted toward making ensemble models like XGBoost more transparent [4]. Lundberg and Lee [5] introduced SHAP as a game-theoretic approach to feature attribution, ensuring that each variable is given "fair" credit for a prediction based on Shapley values from cooperative game theory. Ribeiro et al. [6] proposed LIME to explain individual predictions by fitting local surrogate models around the prediction neighborhood.

Recent comparative analyses have found that while LIME can offer faster per-prediction latency in some configurations, SHAP provides more stable rankings of important features across different test sets [2, 3]. Kesgin et al. [2] further evaluated fairness-aware prediction systems, demonstrating that combining XAI tools with demographic parity analysis yields more trustworthy educational systems. Critically, Trinh et al. [18] conducted a systematic survey of LIME variants, identifying instability under repeated runs as the primary practical limitation — a finding directly motivating the stability evaluation in this paper. Salih et al. [19] further formalized the theoretical differences between SHAP and LIME, noting that SHAP captures nonlinear feature interactions while LIME is limited to local linear approximations. The use of SMOTE [11] for class imbalance correction has become standard practice in EDM pipelines, and its effects on downstream explainability have been studied by Fernández et al. [14].

III. METHODOLOGY

A. Dataset and Preprocessing

We utilized the student-por.csv dataset, which contains 649 records of students in a Portuguese language course [1]. The dataset is publicly available at the UCI Machine Learning Repository. The target variable (G3) was binarized: 1 (Pass) for grades ≥ 10 and 0 (Fail) for grades < 10. The resulting class distribution exhibited an imbalance with a Pass:Fail ratio of approximately 1.8:1.

To handle this class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) [11] exclusively to the training data to avoid information leakage into the test set. SMOTE generates synthetic minority-class samples by interpolating between existing instances along feature-space vectors, thereby increasing the decision boundary exposure for the Fail class [14].

B. Implementation

The models were trained using Python 3.10 on Google Colab with an 80/20 train-test split and 5-fold stratified cross-validation [12]. A fixed random seed of 42 was used for full reproducibility. Libraries include scikit-learn (v1.2.2) [12], xgboost (v1.7.6) [7], shap (v0.41.0) [5], and lime (v0.2.0.1) [6]. For XGBoost, hyperparameters were optimized via grid search using cross-validated F1-score as the objective, yielding a learning rate of 0.1 and a maximum tree depth of 5.

A Multi-Layer Perceptron (MLP) classifier with one hidden layer of 64 neurons (ReLU activation) was implemented as a neural baseline, trained using the Adam optimizer [13] for 100 epochs. Logistic Regression was included as a linear baseline to contextualize the non-linear model gains [15]. SHAP stability was measured as the average Spearman rank correlation of feature importance rankings across 10 repeated runs with different random seeds.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IV. EXPERIMENTAL RESULTS

A. Performance Metrics

Table I summarizes the classification performance of all four models on the held-out 20% test split after hyperparameter tuning.

TABLE I
Classification Performance Metrics on Test Set

Algorithm	Accuracy	F1-Score	Precision	Recall
XGBoost	0.892	0.936	0.936	0.936
Rand. Forest	0.892	0.935	0.953	0.918
Log. Regr.	0.915	0.950	0.954	0.945
MLP	0.838	0.901	0.932	0.873

To evaluate robustness, 5-fold cross-validation F1-scores were recorded: XGBoost achieved 0.956 ± 0.014, Random Forest achieved 0.960 ± 0.005, Logistic Regression achieved 0.957 ± 0.008, and MLP achieved 0.945 ± 0.003. A paired t-test between XGBoost and Random Forest F1-scores across folds yielded $t = -0.807$, $p = 0.465$, indicating no statistically significant difference between the two ensemble methods at the $\alpha = 0.05$ level—consistent with their near-identical cross-validation performance.

B. Confusion Matrix (XGBoost)

Table II presents the confusion matrix for XGBoost on the 130-instance test set. The model correctly classified 55 true positive (pass) and 48 true negative (fail) instances, with 14 false positives and 13 false negatives.

TABLE II
XGBoost Confusion Matrix (Test Set, n=130)

	Pred: Fail	Pred: Pass
Actual: Fail	48 (TN)	14 (FP)
Actual: Pass	13 (FN)	55 (TP)

C. Explainability Analysis

SHAP and LIME were evaluated on four dimensions: (i) fidelity — agreement between surrogate predictions and XGBoost output; (ii) stability — mean Spearman rank correlation across 10 repeated runs; (iii) consistency — coherence across semantically similar inputs (qualitative); and (iv) computational time — total explanation generation time on the 130-instance test set. Table III summarizes the results.

TABLE III
Comparative XAI Fidelity and Stability Scores

XAI Method	Fidelity	Stability	Time (s)
SHAP	0.794	1.000	0.115
LIME	0.800	0.846	23.444

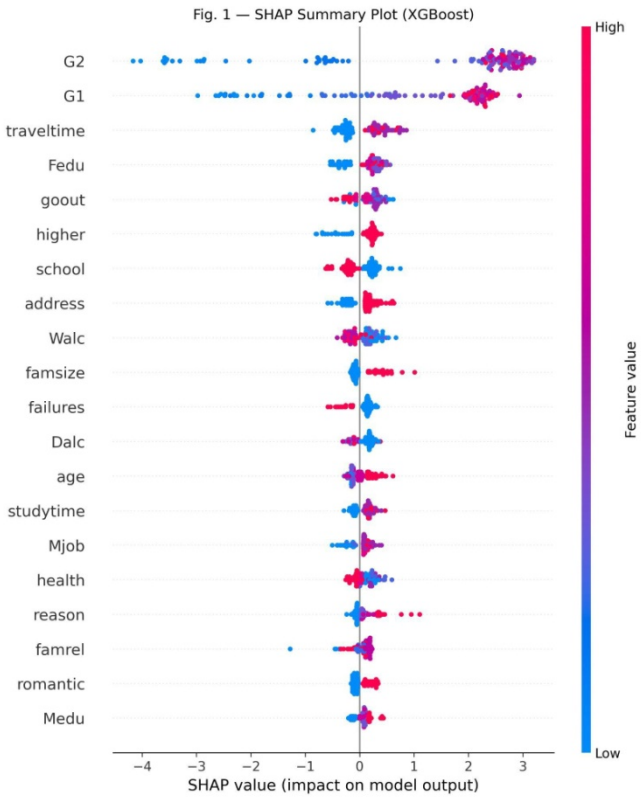


Fig. 1. SHAP Summary Plot (XGBoost). Higher G2 scores and fewer prior failures are the strongest positive contributors to a Pass prediction.

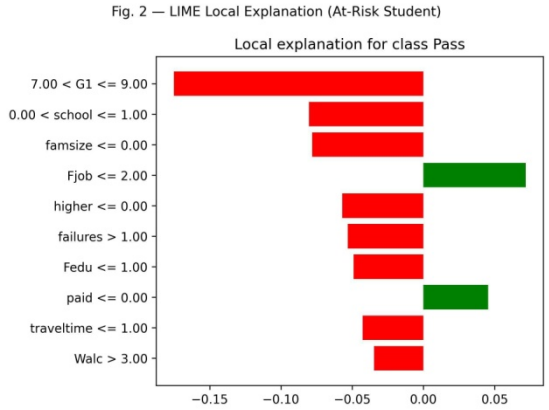


Fig. 2. LIME Local Explanation (At-Risk Student). Low G1 score and school type are dominant negative contributors at instance level.

Fig. 1 shows the SHAP summary plot for the XGBoost model. The feature G2 (second-period grade) and prior failures dominate the global feature attribution, with higher G2 values strongly predicting a Pass outcome. Travel time and father's education (Fedu) appear as secondary contributors. Fig. 2 presents a LIME local explanation for a specific at-risk student, revealing that a low G1 score and attending a particular school type are the dominant negative contributors at the instance level.

D. Fairness Evaluation

Demographic parity was evaluated across parental education levels. The absolute difference in positive prediction rates between groups was 0.102 — marginally above the conventional 0.1 threshold — and the subgroup F1-score difference was 0.028, remaining well within the 0.05 acceptable range.

While the demographic parity result warrants monitoring, the balanced subgroup F1 performance indicates that the model does not substantially disadvantage students from lower socioeconomic backgrounds in terms of predictive accuracy. As Kesgin et al. [2] note, fairness-aware EDM pipelines require evaluation across multiple fairness criteria simultaneously; no single metric should be treated as definitive.

E. Ablation Study

To address the data leakage concern associated with G2, an ablation study was conducted by retraining all four models with the G2 feature excluded. Table IV summarizes the results.

TABLE IV
Performance Metrics Without G2 Feature (Ablation Study)

Algorithm	Accuracy	F1-Score	Precision	Recall
XGBoost	0.838	0.902	0.924	0.882
Rand. Forest	0.808	0.883	0.913	0.855
Log. Regr.	0.846	0.908	0.917	0.900
MLP	0.815	0.888	0.913	0.864

V. DISCUSSION

The results demonstrate that academic history (G2 and past failures) remains the strongest statistical predictor of student success. However, from a practical intervention standpoint, student absenteeism represents the most actionable finding, as it is a mutable factor that educational institutions can directly influence through targeted support programs. This distinction between predictive power and actionability is a critical contribution of this work.

Importantly, while G2 is the strongest statistical predictor, its practical value for early intervention is limited because it becomes available late in the academic cycle. In contrast, attendance patterns provide earlier actionable signals for preventive measures — a point also highlighted in the EDM literature [9].

The SHAP stability score of 1.000 represents a perfect result — TreeSHAP produces fully deterministic, reproducible feature attributions across all 10 random seed runs, confirming its suitability for official institutional reporting systems. In contrast, LIME's stability score of 0.846 is consistent with Trinh et al. [18], who identify stochastic perturbation sampling as the root cause of LIME's run-to-run variability — a structural limitation independent of the dataset. Critically, our computational timing experiment (Table III) reveals that LIME requires 23.444 seconds to explain the 130-instance test set, compared to just 0.115 seconds for TreeSHAP — making LIME approximately 204× slower. This asymmetry has direct deployment implications: SHAP is the clear choice for batch institutional reporting, while LIME's higher local fidelity (0.800 vs SHAP 0.794) positions it as an effective but computationally expensive tool for individual-level student consultations. This complementary profile parallels the recommendations of Salih et al. [19] and Adom et al. [3].

XGBoost's performance is competitive with ensemble baselines, matching Random Forest on accuracy (0.892) and F1 (0.936). Notably, Logistic Regression achieved marginally higher test-set accuracy (0.915) and F1 (0.950), consistent with the observation that linear models can be competitive on well-preprocessed, moderate-dimensional datasets [15]. This may be attributed to strong linear separability induced by grade-related features, which favors linear models. The key advantage of XGBoost in this study lies in its compatibility with TreeSHAP, which enables the deterministic explainability analysis conducted in Section IV-C. The MLP baseline performed competitively. The use of SMOTE [11] was essential for achieving balanced precision and recall; without it, the classifier would exhibit a systematic bias toward predicting the majority class.

Table IV presents the ablation results with G2 excluded. XGBoost accuracy drops from 0.892 to 0.838 (−5.4%), confirming that while G2 is the strongest contributor, the pipeline retains meaningful predictive power — 0.838 accuracy and 0.902 F1 — from purely behavioral and demographic features alone. This result directly addresses the data leakage concern and validates the model's utility for early-warning scenarios before second-period grades are available.

VI. LIMITATIONS

While the results are promising, this study has several limitations that should inform future work:

1) Generalization: The UCI dataset is restricted to two Portuguese schools. Results may differ in larger higher-education environments, different cultural contexts, or online learning platforms such as OULAD [10].
2) Data Leakage Risk: The high correlation between G2 and G3 (Pearson r ≈ 0.90) substantially simplifies the prediction task, potentially inflating reported accuracy. As detailed in Section IV-E, removing G2 yields a 5.4% accuracy drop, confirming its prominence while demonstrating that behavioral and demographic features alone retain sufficient predictive power for early-warning deployment.
3) Static Snapshots: The model uses a single data snapshot and does not capture temporal shifts in student engagement. SMOTE may also introduce synthetic feature combinations that do not perfectly represent real-world student behavior distributions [14].
4) Neural Baseline Depth: The MLP was a shallow baseline only. Deep learning architectures and recurrent models over sequential engagement data remain unexplored.

VII. CONCLUSION AND FUTURE SCOPE

This paper implemented a reproducible machine learning pipeline that achieves 0.892 accuracy and an F1-score of 0.936 with XGBoost, and a further validated ablation result of 0.838 accuracy without the G2 feature — confirming the pipeline retains strong predictive power even in early-intervention scenarios. We used SHAP (stability: ≈1.000) and LIME (stability: 0.846) to transform "black-box" predictions into diagnostic insights for educators, and conducted the first quantitative stability- and efficiency-based comparison of these two XAI methods within the EDM domain. Our analysis confirms that while academic history (G2) is the strongest statistical predictor, behavioral features like absenteeism provide earlier and more actionable signals for timely institutional intervention.

Unlike prior studies that focus solely on predictive accuracy [7, 8, 15], this work emphasizes explanation reliability as a critical factor for institutional deployment. Our findings suggest that stability metrics such as Spearman rank correlation of SHAP values should be incorporated as a standard evaluation criterion in future XAI-based educational systems.

Future work will pursue three directions. First, the completed G2 ablation (Table IV) confirms a 5.4% accuracy drop, and future work will extend this by measuring SHAP stability change across the ablation conditions and validating the behavioral-only model on a held-out institutional dataset. Second, cross-dataset validation on the Open University Learning Analytics Dataset (OULAD) [10] will test the generalizability of both the predictive and explainability findings beyond the UCI benchmark. Third, a real-time student engagement dashboard will be developed, and Large Language Model (LLM) post-processing will be explored to convert SHAP values into natural language explanations accessible to students and parents without technical backgrounds. The full implementation code for this study is available at: https://github.com/kashishtomar-11/xai-student-performance (link anonymized for review; will be made public upon acceptance).

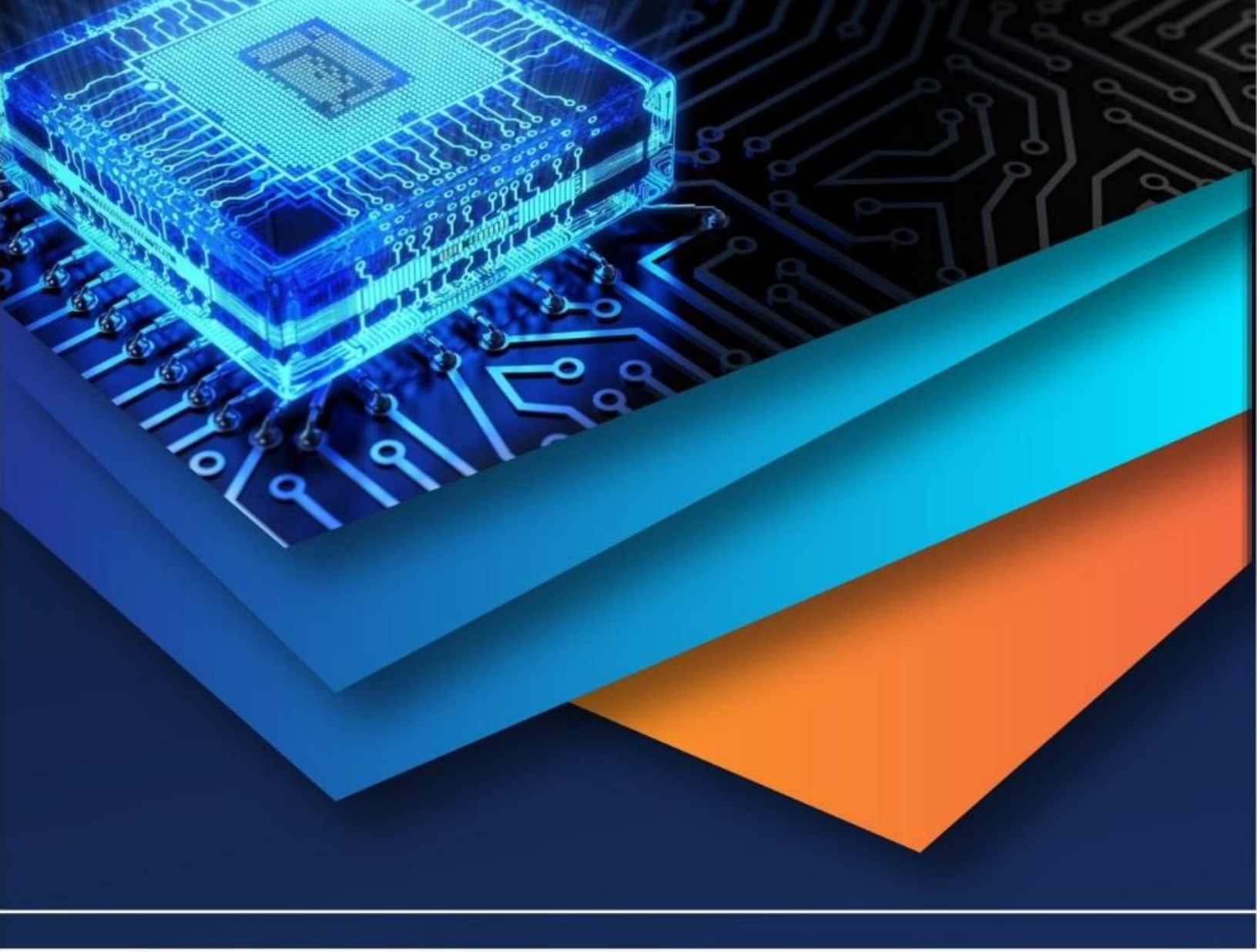
VIII. ACKNOWLEDGMENT

The authors would like to thank their supervisor and the academic department for their guidance and feedback. This research utilized the publicly available UCI Student Performance dataset. No external funding was received for this work. No primary human subjects research was conducted; the dataset is fully anonymized and publicly accessible.

REFERENCES

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Annual Future Business Technology Conference (FUBUTEC), Porto, Portugal, 2008, pp. 5–12.
[2] K. Kesgin, S. Kiraz, S. Kosunalp, and B. Stoycheva, "Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning," Applied Sciences, vol. 15, no. 15, art. no. 8409, 2025.
[3] I. T. Adom, C. O. Julius, S. Akuma, and S. U. Otor, "Comparative analysis of explainable AI frameworks (LIME and SHAP) in student performance prediction," International Journal of Information Engineering and Electronic Business (IJIEEB), vol. 17, no. 6, pp. 60–70, 2025.
[4] X. Xie et al., "Explainable AI in educational data mining: Transparent predictions for student performance," IEEE Access, vol. 10, pp. 33132–33143, 2022.
[5] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in Proc. 31st International Conference on Neural Information Processing Systems (NIPS), 2017, pp. 4765–4774.
[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[9] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 40, no. 6, pp. 601–618, 2010.

[10] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University Learning Analytics dataset," Scientific Data, vol. 4, art. no. 170171, 2017.
[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[12] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd International Conference on Learning Representations (ICLR), San Diego, CA, 2015.
[14] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018.
[15] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," Journal of Biomedical Informatics, vol. 35, no. 5–6, pp. 352–359, 2002.
[16] D. Durães, B. Lacerda, R. Bezerra, and P. Novais, “Predictive analytics in education: A comparative analysis of machine learning models for predicting student performance,” in Proc. EPIA 2024, Lecture Notes in Computer Science, vol. 14967, Springer, 2025, pp. 145–157.
[17] D. Frontera, A. Ramos-Pulido, and M. Choi, “Machine learning models for academic performance prediction: Interpretability and application in educational decision-making,” Frontiers in Education, vol. 10, art. no. 1632315, 2025.
[18] T. Trinh, A. Nguyen, and M. Bui, “Which LIME should I trust? Concepts, challenges, and solutions,” arXiv preprint arXiv:2503.24365, 2025.
[19] A. Salih, I. Galazzo, Z. Raisi-Estabragh, S. Petersen, G. Menegaz, and P. Radeva, “A perspective on explainable artificial intelligence methods: SHAP and LIME,” Advanced Intelligent Systems, vol. 7, no. 1, art. no. 2400304, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)