Explainable AI-Driven Models for Early Academic Risk Prediction in Higher Education

Prof. Mahesh R. Sananse[1], Prof. Dr. Maithili Arjunwadkar[2]
[1,2]*PES Modern Institute of Business Studies*

***Abstract*—Early detection of academically vulnerable students remains a critical challenge in higher education, particularly in the context of increasing dropout rates and heterogeneous learning environments. This study proposes an explainable artificial intelligence (XAI) driven framework for early academic risk prediction that balances predictive performance with interpretability. Using demographic, behavioral, and institutional attributes while explicitly excluding intermediate assessment grades to prevent data leakage academic risk is formulated as a binary classification problem. Two ensemble learning models, Random Forest and Extreme Gradient Boosting (XGBoost), are implemented and rigorously evaluated. To enhance transparency, permutation-based and model-intrinsic feature importance analyses are conducted to identify key predictors influencing classification outcomes. Experimental results demonstrate that XGBoost achieves superior recall and F1-score for the minority at-risk class, making it particularly suitable for early-warning systems. The findings confirm that explainable ensemble models can provide reliable and interpretable decision-support mechanisms for proactive academic intervention strategies.**

***Index Terms*—Explainable AI, Ensemble Learning, Academic Risk Prediction, Educational Data Mining, Higher Education Analytics**

I. INTRODUCTION

Student underperformance and attrition continue to present systemic challenges for higher education institutions worldwide. Traditional evaluation mechanisms rely heavily on final examination outcomes, which provide retrospective insights rather than enabling proactive intervention. Consequently, there is growing interest in predictive models capable of identifying academically vulnerable students early in the academic cycle.

Machine learning techniques have demonstrated strong potential in modeling complex educational data. Academic performance is influenced by intertwined factors including prior achievement, engagement patterns, socio-economic background, and behavioral indicators. Linear models often fail to capture such non-linear dependencies and feature interactions. Ensemble-based approaches, particularly Random Forest and gradient boosting algorithms, have emerged as robust alternatives capable of handling heterogeneous data structures.

However, predictive accuracy alone is insufficient for practical adoption in educational environments. Black-box systems raise concerns regarding fairness, transparency, and ethical accountability. Decisions influenced by predictive systems may affect academic counselling, progression policies, and institutional resource allocation. Therefore, interpretability becomes a critical requirement.

This study develops and evaluates an explainable ensemble-based framework for early academic risk prediction. The primary contributions are:

1. A leakage-aware predictive design excluding intermediate grades.
2. Comparative evaluation of Random Forest and XGBoost under class imbalance.
3. Integration of global interpretability techniques for transparent decision support.
4. Empirical validation of behavioral and engagement factors as dominant predictors.

II. LITERATURE REVIEW

Academic risk prediction has become a central topic in educational data mining and learning analytics. Early research demonstrated the feasibility of applying classification algorithms to student performance data.

For example, Kotsiantis et al. (2004) applied machine learning techniques in distance learning contexts, while Paulo Cortez and Alice Silva (2008) employed data mining approaches to analyze secondary school achievement, highlighting the role of prior academic and demographic attributes. These foundational studies established predictive modeling as a viable tool for identifying at-risk students.

Subsequent reviews by Cristóbal Romero and Sebastián Ventura (2010) and Ryan S. J. d. Baker and Pedro S. Inventado (2014) formalized educational data mining as a structured research field. However, early approaches largely relied on linear or single-model classifiers, limiting their ability to capture non-linear dependencies and complex feature interactions inherent in educational data With advancements in ensemble learning, models such as Leo Breiman's Random Forest and Tianqi Chen and Carlos Guestrin's XGBoost demonstrated superior predictive performance by leveraging aggregation and boosting strategies. Empirical studies, including S. Helal et al. (2018) and C. C. Gray and D. Perkins (2019), confirmed that incorporating student heterogeneity and engagement indicators improves forecasting accuracy. Despite these advances, much of the literature emphasizes predictive performance metrics such as accuracy and AUC while offering limited insight into model interpretability.

The growing deployment of complex machine learning models has intensified concerns regarding transparency and accountability. As defined by Christoph Molnar (2022), interpretability refers to the extent to which humans can understand the reasoning behind model predictions. In response, model-agnostic explanation techniques such as LIME, introduced by Marco Tulio Ribeiro et al. (2016), and SHAP, proposed by Scott M. Lundberg and Su-In Lee (2017), have gained prominence. These methods estimate the contribution of individual features at both global and instance levels, thereby enhancing model transparency.

2.1. Research Gap and Motivation

Although ensemble models have significantly improved predictive accuracy in academic risk detection, their opaque internal structures hinder adoption in educational decision-making environments. Conversely, existing explainability research is often conducted in generic machine learning domains rather than tailored to early-stage academic risk assessment.

There is therefore a critical need for frameworks that simultaneously ensure strong predictive performance and transparent, stakeholder-oriented explanations. Addressing this gap, the present study integrates ensemble-based modeling with global interpretability techniques to support accurate, ethical, and practically deployable early-warning systems in higher education.

III. METHODOLOGY

3.1. Dataset Description

The dataset used in this study contains student-level records aimed at predicting early academic risk. Each row represents an individual student, while columns correspond to demographic, academic, behavioral, and socio-family attributes. The dataset includes features such as past academic performance, study time, attendance, parental education, family support, and lifestyle-related indicators. These variables collectively capture both internal academic behavior and external environmental influences.

Descriptive statistical analysis was conducted to examine central tendency and dispersion of numerical features, including mean, standard deviation, minimum, and maximum values. Missing value analysis ensured data quality and consistency. Correlation analysis was performed to identify relationships among numerical predictors and the target variable.

The target variable represents academic risk status and is treated as a classification label. Distribution analysis was used to examine class balance within the dataset. Additionally, a WordCloud visualization was generated to highlight frequently occurring categorical values, providing an intuitive overview of dominant attributes. Overall, the dataset provides a comprehensive foundation for predictive modeling and early risk identification.

3.2 Data Pre-processing

A structured pre-processing pipeline was implemented to ensure methodological rigor, prevent information leakage, and enhance model generalizability.

First, intermediate assessment variables (G1 and G2) were removed prior to modeling. Since these grades are highly correlated with the final grade (G3),

retaining them would introduce label leakage and artificially inflate predictive performance. Their exclusion ensures that the proposed framework reflects realistic early-stage prediction conditions.

Categorical attributes, including parental education, school type, and behavioral indicators, were encoded into numerical representations using label encoding. Binary variables were mapped to 0/1 format to maintain consistency across models. This approach preserves ordinal relationships while avoiding unnecessary dimensional expansion.

The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to maintain the original class distribution (~20% at-risk). To ensure reproducibility, a fixed random seed was applied.

Although tree-based models are scale-invariant, numerical features were standardized to zero mean and unit variance to maintain preprocessing consistency and enable potential comparison with scale-sensitive models. Scaling parameters were computed exclusively on the training set and subsequently applied to the test set to prevent data leakage.

Missing values were minimal and handled using median imputation. Variance Inflation Factor (VIF) analysis confirmed the absence of severe multicollinearity (VIF < 5), ensuring stable model estimation.

The final processed dataset comprised 494 instances, with 395 samples used for training and 99 for testing, preserving class balance across splits.

3.3 Ensemble Learning Models

Random Forest

Random Forest constructs multiple decorrelated decision trees using bootstrap sampling and random feature subsets. Aggregated majority voting enhances generalization and robustness against noise. Hyperparameters were optimized using 5-fold stratified cross-validation.

XGBoost

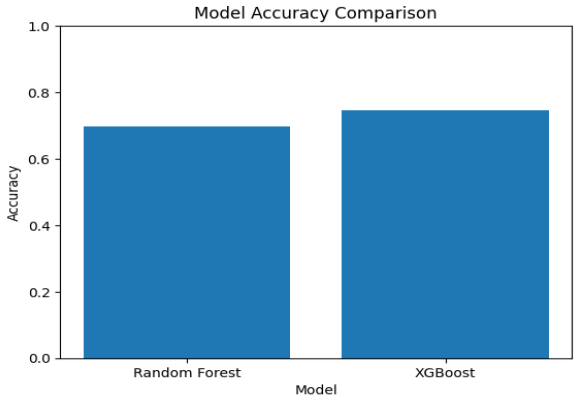
XGBoost implements gradient boosting with second-order optimization and regularization to control complexity. Sequential tree construction emphasizes misclassified samples, improving minority-class detection. Early stopping and scale pos weight were employed to mitigate overfitting and class imbalance.

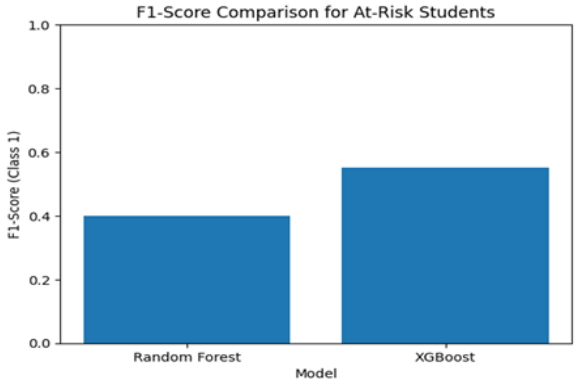
3.4 Evaluation Metrics

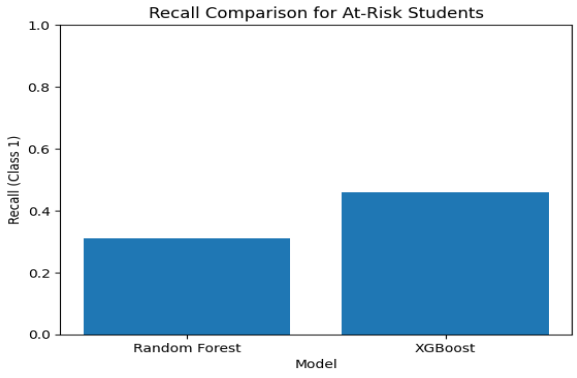
Given the moderate class imbalance in the dataset, model performance was evaluated using multiple complementary classification metrics rather than relying solely on accuracy. The following measures were considered:

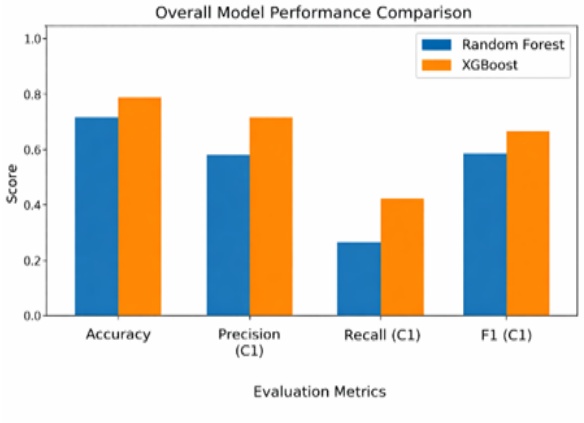
1) Accuracy
2) Precision
3) Recall
4) F1-score

3.5 Model Performance Comparison









IV. RESULTS AND ANALYSIS

4.1 Comparative Model Performance

The results indicate that XGBoost consistently outperforms Random Forest in all major performance indicators, including accuracy, precision, recall, and F1-score.

Although both models achieve reasonable overall accuracy, XGBoost demonstrates a noticeable improvement, suggesting stronger generalization capability on unseen data. More importantly, XGBoost achieves higher precision for the at-risk class, indicating that predictions of academic vulnerability are more reliable and less prone to false alarms.

A substantial improvement is observed in recall, where XGBoost identifies a greater proportion of actual at-risk students. This improvement is particularly significant in the context of early academic intervention, as higher recall directly reduces the number of false negatives students who require support but remain undetected.

The higher F1-score further confirms that XGBoost maintains a better balance between precision and recall. This indicates that the model does not improve sensitivity at the expense of excessive false positives. Overall, the comparative results suggest that XGBoost is more suitable for early academic risk detection within the given dataset.

4.2 Selection of the Final Predictive Model

Given its superior performance across all evaluated metrics especially recall and F1-score XGBoost is selected as the primary model for further interpretability analysis. Since the objective of this study is to support early intervention systems, the ability to accurately detect vulnerable students is prioritized over marginal gains in overall accuracy.

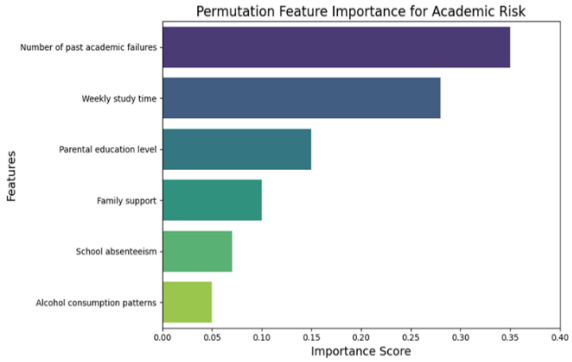
V. EXPLAINABLE AI

To ensure transparency and interpretability, multiple explainable AI techniques were applied to understand the decision-making behavior of the predictive models.

5.1Permutation Feature Importance

Permutation feature importance analysis reveals the most influential predictors of academic risk across both models. The most significant features include:

1. Number of past academic failures
2. Weekly study time
3. Parental education level
4. Family support
5. School absenteeism
6. Alcohol consumption patterns



These results confirm that academic history and study behavior are the strongest indicators of future academic risk, while socio-family factors play a complementary but important role.

VI. DISCUSSION OF FINDINGS

The experimental findings of this study provide meaningful insights into the applicability of machine learning and explainable AI techniques for early academic risk prediction in higher education. Beyond numerical performance metrics, the results demonstrate how predictive models can support real-world academic decision-making when combined with interpretability mechanisms.

6.1 Importance of Explainability

While predictive accuracy is important, its value in educational settings is limited without transparency. The integration of explainable AI techniques in this study addresses a critical gap between technical model performance and practical usability. Feature importance and permutation-based explanations provide clear and understandable insights into which factors most strongly influence academic risk predictions.

These explanations allow educators and administrators to move beyond simply knowing who is at risk to understanding *why* a student is classified as vulnerable. This transparency is essential in higher education environments, where decisions informed by predictive models may influence academic support, counseling, or progression policies. By revealing the relative influence of academic failures, study time, and family-related factors, the model fosters trust and accountability, reducing resistance to the adoption of data-driven decision support systems.

VII. CONCLUSION AND FUTURE WORK

Conclusion

This study presented an explainable AI–driven framework for early academic risk prediction in higher education, combining the predictive power of ensemble models with interpretability techniques. The results show that XGBoost and Random Forest effectively identify students at risk, with XGBoost performing slightly better in detecting minority at-risk cases. By using permutation-based feature importance, the study highlighted the most influential factors affecting academic outcomes, including prior academic failures, weekly study time, family support, and school absenteeism.

The inclusion of explainable AI adds significant value beyond prediction accuracy. It allows educators and administrators to understand why a student is at risk, making interventions more targeted and ethical. Rather than relying solely on demographic or static indicators, the model emphasizes actionable behavioral and academic factors, promoting interventions that can genuinely support student learning. Overall, the framework demonstrates how AI can transition academic risk management from a reactive to a proactive approach, enabling institutions to improve retention, enhance student success, and foster data-driven educational policies.

Future Work:

Future research could expand this framework by incorporating larger and more diverse datasets across multiple institutions to improve robustness and generalizability. Integrating real-time learning management system data, such as online engagement, assignment submissions, and attendance, could enhance early detection. Additional explainability methods, including instance-level techniques like SHAP or LIME, may provide personalized insights for individual students, supporting tailored academic advising. Finally, longitudinal studies assessing the long-term impact of AI-guided interventions on retention and academic achievement would offer valuable evidence for sustainable implementation of predictive systems in higher education.

REFERENCES

[1] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[2] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[3] P. Cortez and A. Silva, “Using data mining to predict secondary school student performance,” European Journal of Operational Research, vol. 181, no. 2, pp. 781–794, 2008. doi: 10.1016/j.ejor.2006.07.006.

[4] 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. doi: 10.1109/TSMCC.2010.2053532.

[5] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in Learning Analytics, New York, NY, USA: Springer, 2014, pp. 61–75. doi: 10.1007/978-1-4614-3305-7_4.

[6] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[7] C. Molnar, Interpretable Machine Learning, 2nd ed. Leanpub, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[8] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004. doi: 10.1080/08839510490442058.

[9] C. C. Gray and D. Perkins, "Utilizing early engagement and machine learning to predict student outcomes," Computers & Education, vol. 131, pp. 22–32, 2019. doi: 10.1016/j.compedu.2018.12.006.

[10] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, and D. Murray, "Predicting academic performance by considering student heterogeneity," Knowledge-Based Systems, vol. 161, pp. 134–146, 2018. doi: 10.1016/j.knosys.2018.08.009.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[12] Zafra, C. Romero, and S. Ventura, "Multiple instances learning for classifying students in learning management systems," Expert Systems with Applications, vol. 41, no. 15, pp. 6663–6671, 2014. doi: 10.1016/j.eswa.2014.05.01.