International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(11s) 2026 pp. 1008-1021
https://doi.org/10.70917/ijcisim-2026-3830
Article

An Explainable Stacking Ensemble Learning Framework for Early Student Attrition Prediction and Risk Stratification Using Educational Data Mining

Rupali Ambalal Jadhav[1], Rupal Parekh[2]

[1]Atmiya University, Rajkot, Gujrat, India, Email- smarupa@gmail.com
[2]Atmiya University , Rajkot, Gujrat, India, Email- rupal.kachalia@atmiyauni.ac.in

Abstract: Student attrition poses a significant challenge in higher education. It negatively impacts academic performance, reputation, and resource management. This paper presents an XAI-based stacking ensemble for early prediction of student attrition with EDM methods.

We used a publicly available dataset on Kaggle, which contained information about 4,424 undergraduate students and their personal characteristics, academic status and socioeconomic background.

Class imbalance is dealt with using a cost-sensitive learning strategy with the parameter of scale_pos_weight combined with stratified train-test splitting. The proposed stacking ensemble framework used the Random Forest and XGBoost as base models and Logistic Regression as a meta-learner. The selected threshold of classification for early prediction was 0.20. The proposed model performance is assessed with several performance metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, as well as calibration analysis.

Experimental results showed that the proposed stacking ensemble outperformed individual base models and offered the best prediction accuracy with a high Recall value in early dropout prediction.

The developed model interpretability by SHAP and LIME showed that student performance and financial situation played a significant role in predicting the risk of attrition at early stages of study. The predicted probabilities are divided into the categories of Low, Medium and High-Risk using percentiles, to facilitate interventions for the students at risk. In this context, we offered an interpretable and applicable artificial intelligence-driven early warning system in higher education.

Keywords: Student Attrition Prediction · Educational Data Mining · Explainable Artificial Intelligence (XAI) · Early Warning System · Learning Analytics

1. Introduction

Student attrition remains one of the most persistent challenges faced by higher education institutions worldwide. Student dropout not only affects learners' academic and professional careers but also has significant consequences for institutional performance, reputation, accreditation outcomes, and effective resource utilization. High attrition rates result in financial losses, reduced graduation rates, inefficient allocation of academic resources, and diminished institutional rankings. Consequently, early identification of students at risk of dropping out has become an important research area in Educational Data Mining (EDM) and Learning Analytics. The proposed research High rates of dropout lead to increased institutional costs, reduced graduation rates, poor deployment of academic resources, lower ranking of institutions, etc. So, the early prediction of at-risk students' dropout is now an increasingly popular research problem in Educational Data Mining (EDM) and Learning Analytics.

The research work, in recent years, has been focusing on developing a machine learning model to predict student dropout. During the last few years, with the advancement of EDM and machine learning, many institutions have analyze student's education data sets of large sizes for the prediction of the performance and drop-out rate of the students. The machine learning models like decision tree (J48), Support Vector Machine (SVM), random forest, gradient boosting machine (GBM), XGBoost, deep learning models and ensemble models (bagging and boosting) have demonstrated high predictive power on the identification of students at risk of dropping out. But, most of these methods focused only on maximising predictive accuracy, disregarding other significant parameters like interpretability, transparency, calibration of the models, and support to decision making. So, most of the developed models often end up as ‘black-box models’ that are difficult to explain for the educators, student advisors, institutional leaders and to gain confidence while implementing it to make plans for intervention strategies.

One of the limitations is that the majority of the prediction models consider solely historical static academic information (data at one specific point in time), and neglect the dynamics of student’s academic journey, such as students engagement, academic progression, learning behaviours and ever-changing risk pattern throughout their time in education. In addition to this, a large number of institutions find out that students have already drastically dropped academically before it’s being flagged up by their warning systems.

Recently, much research has focused on utilizing explainable artificial intelligence (XAI) to develop more explainable and trust-worthy educational prediction systems. With XAI techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), educators, academic advisors and decision makers could get insight into how individual features influence the model’s predictions and provide effective support for students with evidence-based and individualized recommendations. Nevertheless, a review of the current literature reveals a research gap in building explainable ensemble machine learning models with a comprehensive early warning system for practical application to identify potential student dropouts and their reasons for dropping out.

Another challenge commonly encountered in student attrition prediction is class imbalance, where the number of continuing students substantially exceeds the number of dropout cases. Conventional resampling techniques may introduce synthetic bias or information loss, potentially affecting model generalization. Cost-sensitive learning strategies provide an effective alternative by assigning greater importance to minority-class samples during model training while preserving the original data distribution.

To address these limitations, this study proposes an Explainable Artificial Intelligence (XAI)-based stacking ensemble framework for early student attrition prediction using Educational Data Mining techniques. The proposed framework employs Random Forest and XGBoost as complementary base learners and Logistic Regression as the meta-learner to improve predictive performance while maintaining model stability. Instead of synthetic oversampling, class imbalance is handled through a cost-sensitive learning strategy using the scale_pos_weight parameter combined with stratified train-test splitting. A lower classification threshold of 0.20 is adopted to improve early identification of at-risk students by prioritizing recall, thereby enabling institutions to initiate timely academic interventions.

The proposed framework is evaluated using a publicly available Kaggle dataset containing records of 4,424 undergraduate students with demographic, socioeconomic, and academic attributes. Model performance is assessed using Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, and calibration analysis to provide a comprehensive evaluation of predictive reliability. Furthermore, SHAP and LIME are employed to explain both global and local model predictions, revealing the key factors influencing student attrition. Based on predicted probabilities, students are categorized into Low-, Medium-, and High-risk groups using percentile-based risk stratification to support prioritized intervention planning.

2. LITERATURE REVIEW AND RESEARCH GAP

The connected graph displays both visually and analytically how the various components of the research project "An Explainable Machine Learning Framework for Early Prediction of Student Attrition Using Educational Data Mining" work together. The nodes in the graph are the research components, which include the ML, EDM, XAI, student data sources, early prediction components, and institutional intervention systems

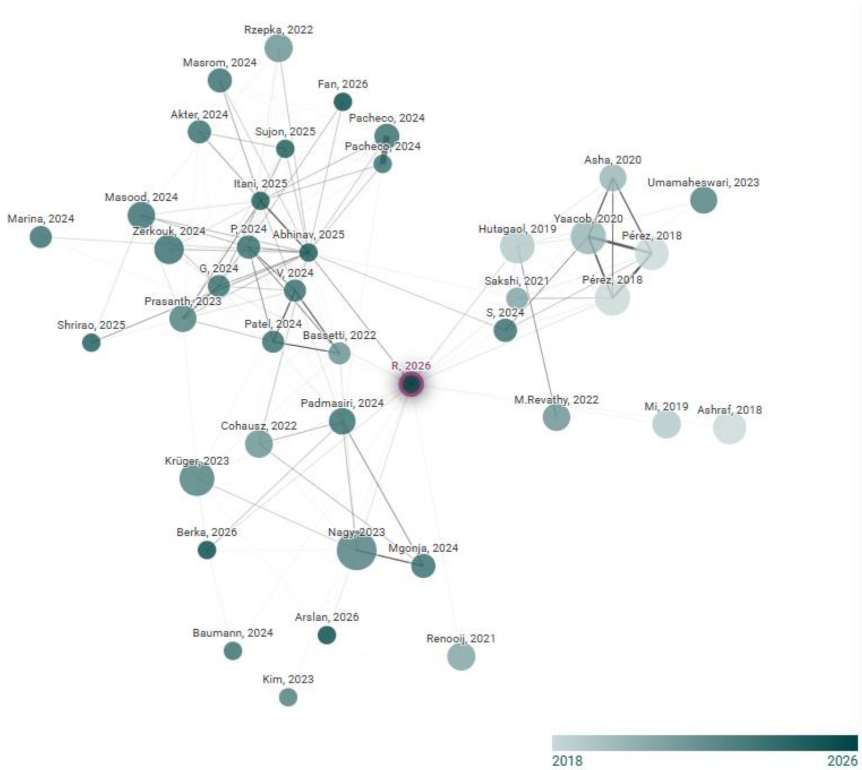

Figure. 1. Literature network visualization generated using Connected Papers, illustrating research clusters in student attrition prediction

The edges indicate how these nodes are interconnected and how they rely on each other. For example, EDM extracts information from various student data sources that are fed into the ML algorithms that ultimately produce a prediction for institutional decision-making. Literature mapping performed through Connected Papers shows that even though much research has been performed related to ML and student attrition prediction[10], that research has predominantly focused on prediction accuracy with little to no concern for interpretability or actionability of research results. Observing both the connected graph and the literature gap table, although each component may have a lot of individual research performed about it, the interaction between the components remains very weak and underdeveloped. A clear disconnection can be identified between ML prediction and explainability[18], as current student attrition prediction models may behave like black boxes, failing to provide understandable feedback that can build stakeholder trust. A second major disconnect can be found between early prediction and real-time behavioral /learning analytics data, as current studies primarily rely on past academic records instead of tracking live activity. The last missing link seems to exist between the prediction component and the institutional decision-support component, as while predictions can be made about student attrition risk, there is a lack of research on transforming these results into a structured approach for institutional intervention. The connected graph assisted by the Connected Papers literature mapping systematically points out underdeveloped or disconnected relationships between the components, and a definite research gap exists where individual components have been sufficiently investigated, but no holistic approaches have been introduced combining explainable ML algorithms and early intervention for student retention[21].

Table 1. Gap Analysis

Sr. No.	Research Component (Graph Node)	Existing Research Focus	Connection with Other Components	Identified Limitation	Research Gap
1	Student Attrition Prediction	Prediction using ML models	Connected to academic, behavioral, demographic data	Focus mainly on prediction accuracy	Lack of interpretable prediction models

2	Educational Data Mining (EDM)	Data extraction and preprocessing	Links data sources to ML algorithms	Limited integration of heterogeneous datasets	Need unified EDM framework
3	Machine Learning Algorithms	Random Forest, SVM, Neural Networks	Connected with prediction outcomes	Black-box decision making	Need Explainable ML approaches
4	Explainable AI (XAI)	Feature importance and visualization	Connected to ML model interpretation	Rarely applied in student attrition studies	Integration of XAI with attrition prediction
5	Early Warning Systems	Identification of at-risk students	Connected to institutional intervention	Prediction occurs too late	Need early-stage prediction models
6	Student Behavioral Data	LMS activity, attendance, engagement	Connected to learning analytics	Under utilized temporal behavior patterns	Incorporate real-time behavioral analytics
7	Institutional Decision Support	Administrative policy making	Linked with prediction results	Results not actionable	Develop interpretable decision-support framework
8	Model Evaluation	Accuracy, Precision, Recall	Connected to model validation	Explainability metrics ignored	Combine performance + explainability evaluation
9	Intervention Strategies	Counseling and mentoring	Connected to prediction output	No automated recommendation system	Explainable intervention recommendation system

3. PROPOSED METHODOLOGY

3.1 System Architecture

The students' data was gathered across three important aspects: demographic, academic and behavioral parameters, which are important in students' attrition prediction. Collected students' data underwent several pre-processing steps, which include data cleaning, missing value imputations, feature encoding, feature scaling and preparing structured data for machine learning models[22,30].

ML algorithms produce accurate predictions, but they act as a black-box model, and do not provide an understanding about their decision made. Educators don't trust such a model, and it restricts the application of predictive analytics in the education setup. To overcome this issue, XAI techniques (SHAP and LIME) were adopted to produce interpretable global and local explanations of ML model predictions [1]

Figure 2 presents the proposed student dropout prediction framework; the framework consists of data collection, data pre-processing, baseline model building, ensemble learning, prediction of risk and explanation by using XAI.

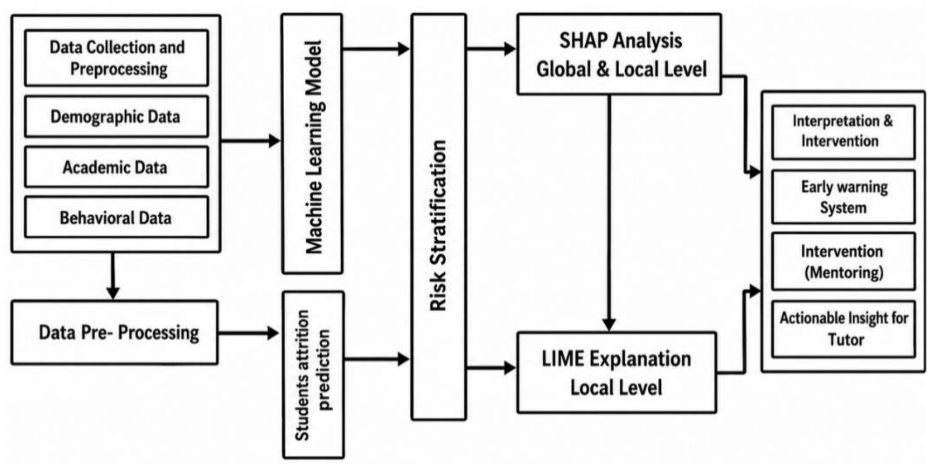

Figure 2. System architecture for an explainable AI-driven student attrition model

3.2 Dataset Description

This study uses a database containing from Kaggle that contains 4,424 records of undergraduate students. The dataset contains many-faceted attributes, ranging from demographic features to socio-economic backgrounds to admission information to results. Relevant attributes are gender, age at admission, nationality, parents' education/occupation, scholarship, fee paid, program attributes, first-semester results and academic outcome. The selected attributes will function as input variables and allow for student behaviour and learning performance analysis. The target attribute, academic status of a student, can be used for the prediction of student retention/dropout and can be of service for a more comprehensive student retention model for Higher Education Institutions

3.3 Handling Class Imbalanced

The suggested model, a three-level approach including data pre-processing, algorithm tuning, and model interpretation, addresses the class imbalance problem effectively. A data-level solution in terms of a stratified train-test split is proposed to uphold the intrinsic class proportion, enabling an unbiased assessment of the training and testing datasets and subsequent model fitting and testing. An algorithm-level solution is achieved by introducing a class-weighting technique in the LightGBM classifier that facilitates learning from the minority-class cases to help in detecting students who are at risk of dropping out. A model interpretability-level solution, such as SHAP analysis, is implemented to demonstrate and elucidate the contribution of each feature to the individual predictions and ensure transparent decision-making. This three-level solution improves model accuracy, fairness and explainability, thus allowing for accurate and reliable detection of at-risk students and timely interventions.

In this experiment, a dataset consisting of 4,424 student records was utilized. The class variable, Target, represents the student attrition. The target variable contains two categories: Graduate, indicating students who completed the program, and Dropout, representing students who discontinued their studies. This target variable is used to train predictive models for identifying students at risk of attrition. In actual academic settings, the number of students who continue is usually significantly larger than the number of students who drop out, which gives rise to a class imbalance problem. When class imbalance is not properly handled, a machine learning model can still yield a high overall accuracy while not accurately predicting the student dropout risks, which is the goal of this study.

Stratified Train-Test Split has been implemented to resolve this issue, which divides the entire dataset into 80% of training data and 20% of testing data, by retaining the identical ratio of dropped students to continued students present in the original dataset. By doing so, both training and testing data subsets will exhibit the same distribution of the class variable, unlike random splitting that might yield disparate distributions.

Class-weighting has also been employed during the model training process as it might still present an impact of class imbalance, even after Stratified Train-Test Split, using the LightGBM classifier. This assigns an importance of higher weight to the minority-class cases, the dropped students. The class imbalance parameter is defined as:

scale_pos_weight=Number of Dropped Students / Number of Continued Students.

This assigns a penalty when the dropped students are misclassified, which enables getting a higher recall on the minority class for students predicted to drop out.

Using class weighting rather than resampling (like oversampling or SMOTE)[13] allows the model to learn from data while not overly biasing towards majority class predictions.

Table 2. Framework for Handling class Imbalanced

Technique	**Purpose**	**Strength**	**Role in Proposed Framework**
Class Imbalance Handling (Class Weighting)	Balances minority and majority class distribution	Improves detection of rare events and reduces bias	Ensures fair learning and improves minority class prediction
Stratified Train–Test Split	Maintains class proportion during data splitting	Prevents sampling bias	Ensures reliable model validation
LightGBM Classifier	Performs efficient classification	High accuracy and fast training	Acts as the main prediction model
SHAP Explainability	Explains model decisions	Transparent feature importance interpretation	Provides interpretable and trustworthy AI outcomes

3.4 Model Implementation

In this experiment, techniques of advanced machine learning were adopted for the student attrition prediction with reduced risk of overfitting, where a model shows good results for training data but not for unseen data. We used a Random Forest (RF) classifier to enhance the stability of prediction by performing bootstrap aggregating, which is able to decrease the variance and minimise the overfitting by averaging multiple decision trees.

Besides, we applied XGBoost as a gradient boosting algorithm with built-in regularisation parameters to restrain the complexity, boost the prediction accuracy and also overcome overfitting.

Finally, by using two algorithms-Random Forest and XGBoost as base learners and Logistic Regression as the meta-learner- a Stacking Classifier was built to overcome bias and variance, as well as enhance the generalisation accuracy, providing more stable predictions of student attrition.

In conclusion, the ensemble techniques of Bagging, Boosting, and Stacking together formed a powerful framework capable of effectively classifying at-risk dropout students and achieving strong performance in generalisation.

3.5 Risk Stratification

The expected probabilities of dropout were grouped into three levels of risk: Low, Medium and High based on given range values for these levels, illustrated in Table 3, to enable effective strategies.

Classification, the context of an educational intervention calls for the use of risk levels rather than a binary class. As such, these probability scores were therefore re-classified into three meaningful risk levels, Low, Medium, and High, to reflect differentiated action by the institution. These categories were derived from the distribution of 33rd and 66th percentiles of predicted probabilities using percentile-based stratification, thereby providing an evidence-based framework for prioritising institutional interventions. The stratification allows for a continuum ranging from proactive monitoring (low risk students), academic assistance to targeted medium risk students, and intervention for high risk students

Table 3. Risk Classification based on Predicted Probability percentiles.

Category	Range

Low	0.000 – 0.079
Medium	0.079 – 0.484
High	0.484 – 1.000

3.6 Explainability in Attrition Prediction

To improve transparency and practical use in educational decision-making, XAI methods were implemented in the proposed model using SHAP and LIME:

SHapley Additive exPlanations (SHAP) was utilized to provide both local and global interpretations. Globally, it illustrates the most significant features of student attrition from the overall dataset (such as the percentage of attendance, mark of the internal evaluation, hour of engagement in the course), and locally it justifies a specific prediction indicating how much each feature contributed to predicting whether a student would drop out or continue. Teachers or institutes could identify the exact reasons why a specific student might drop out:

Local Interpretable Model-Agnostic Explanations (LIME) was utilized to offer explanations for an individual prediction instance; this involves building a local interpretable approximation of the complex model with a simple and easily understandable model and revealing what features the single prediction is highly dependent on, providing valuable insights to the teachers and administrators.

SHAP and LIME would explain how single features affect the prediction of the student attrition prediction model, to promote transparent model output and the user's trust towards the model, enabling quick action to provide interventions[18,32,33]. It ensures fairness, accountability, and efficiency of decision-making through AI analytics in education. This turns attrition prediction from a predictive application into a decision-support system, allowing timely intervention and a fact-based retention strategy.

4. EXPERIMENTAL RESULTS

4.1 Class distribution Analysis

The dataset used in the study includes a total of 4,424 student records, with student attrition status as the target variable. The class distribution of the dataset is 67.88% for Graduate students and 32.12% for Dropout students. There is a moderate level of class imbalance.

Unequal distributions between classes can result in a machine learning model learning the attributes that classify a majority class, but it won't learn attributes that help classify students at risk for withdrawal. Thus, a cost-sensitive learning technique by tuning the scale_pos_weight parameter was used in XGBoost classifier. scale_pos_weight was set based on the class imbalance ratio to adjust the cost function in accordance with the following calculation of the weight and the original number of positive cases in the training data (i.e., Dropout students) for scaling the loss function associated with the positive class: . Adjusted scale_pos_weight 3.17. Thus, scale_pos_weight increased the penalty for misclassifying dropout students, improving model sensitivity towards the minority class while retaining the distribution of training data. The predicted probability scores produced by the developed model were used to achieve the risk-based stratification for the students on the basis of the probability values rather than arbitrary intervals. The 33rd and 66th percentile (i.e.

0.079 and 0.484) of the predicted probability distributions were determined and taken as threshold points.

The probabilities predicted by the developed model for the students in the test dataset were divided into three risk groups, including the Low Risk Group (0.000–0.079), the Medium Risk Group (0.079–0.484) and the High Risk Group (0.484–1.000). According to the results obtained from applying the developed risk stratification frame to the test data, it was found that there were 339 (38.31%) students in the Low Risk Group, 275 (31.07%) in the Medium Risk Group and 271 (30.62%) in the High Risk Group. There is an approximately equal distribution of the test students into the three risk categories using a percentile-based stratification technique.

Each risk category can be interpreted with the predicted likelihood of student attrition, so that a different response strategy can be applied to these student categories, e.g., students with Low Risk can be continuously monitored, students with Medium Risk may benefit from preventive mentoring, and students with High Risk would require urgent intervention and support from the academic counsellors.

Table 4. Probability-Based Risk Stratification

Risk Category	Probability Range	Number of Students	Percentage (%)
Low Risk	0.000 – 0.079	339	38.31
Medium Risk	0.079 – 0.484	275	31.07
High Risk	0.484 – 1.000	271	30.62
Total	—	885	100.00

4.2 Model Performance

Student attrition prediction was carried out using three ensemble learning methods, namely Random Forest (RF), XGBoost (XGB), and the Stacking Classifier (pro-posed, consisting of RF + XGBoost as base learners with Logistic Regression as the meta-learner). Model performance was assessed using the following evaluation metrics: Accuracy, Precision, Recall, F1-score, PR-AUC and ROC-AUC. A threshold value of 0.20 was applied to facilitate the early identification of at-risk students, giving preference to recall.

The Random Forest classifier recorded an accuracy of 0.802, with values of 0.930, 0.630, 0.751, 0.903, and 0.933 for Recall, Precision, F1-score, PR-AUC and ROC-AUC, respectively. The high Recall of 0.930 indicate the good capability of RF model to correctly classify students at-risk of dropping out. This could be highly useful in an early warning system since minimizing false-negative predictions is a primary goal in many educational systems.

The XGBoost model achieved values of 0.791, 0.912, 0.618, 0.737, 0.910, and 0.934 for Accuracy, Recall, Precision, F1-score, PR-AUC and ROC-AUC, respectively.

Overall accuracy and F1-score of XGBoost model were lower than those achieved by Random Forest. The ROC-AUC is comparable with that of the Random Forest model and has the best PR-AUC among the individual base learners. This reflects the gradient boosting mechanism and its inherent regularization capabilities that efficiently capture the complex nonlinear relationship within the data.

Stacking Classifier (RF + XGBoost + LR) provided the best prediction performance across most metrics. The combined prediction of Random Forest and XGBoost via the Logistic Regression meta-learner obtained an accuracy of 0.852, 0.717, 0.891, 0.794, 0.909, and 0.936 for Accuracy, Precision, Recall, F1-score, PR-AUC and ROC-AUC, respectively.

The Stacking Classifier achieved a Recall of 0.891, which is only marginally lower than that achieved by Random Forest model (0.930), and a much improved Precision (0.717) and F1-score (0.794) compared to Random Forest and XGBoost, respectively.

The ROC-AUC value of the proposed stacking classifier is 0.936, which means it outperformed the other base models by accurately discriminating between graduate and dropout students. Thus, the combination has effectively balanced the model and reduces the overall bias and variance for prediction, leading to an increased generalized prediction with a lower false positive rate.

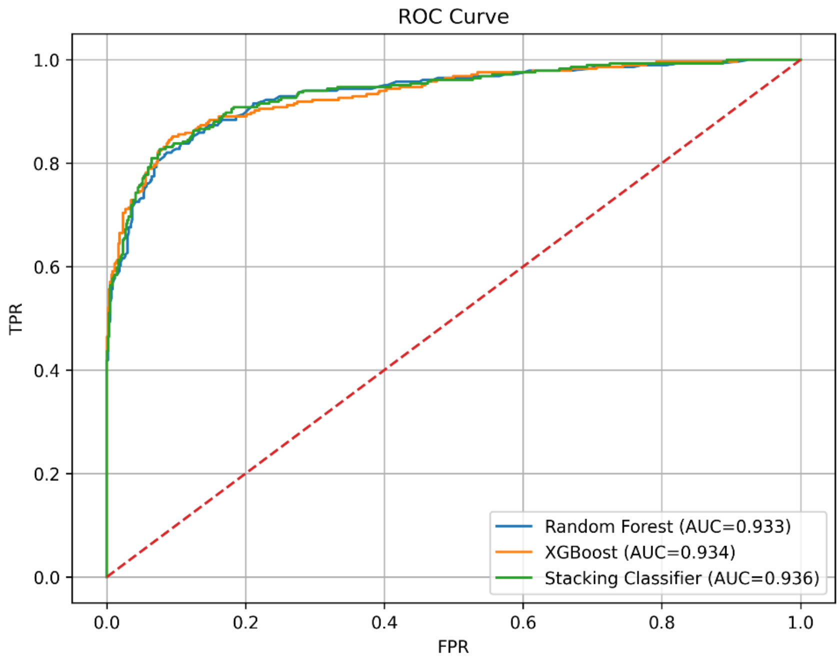

Figure 3. Receiver Operating Characteristic (ROC) Curve Comparison of Ensemble Mod-els for Student Attrition Prediction

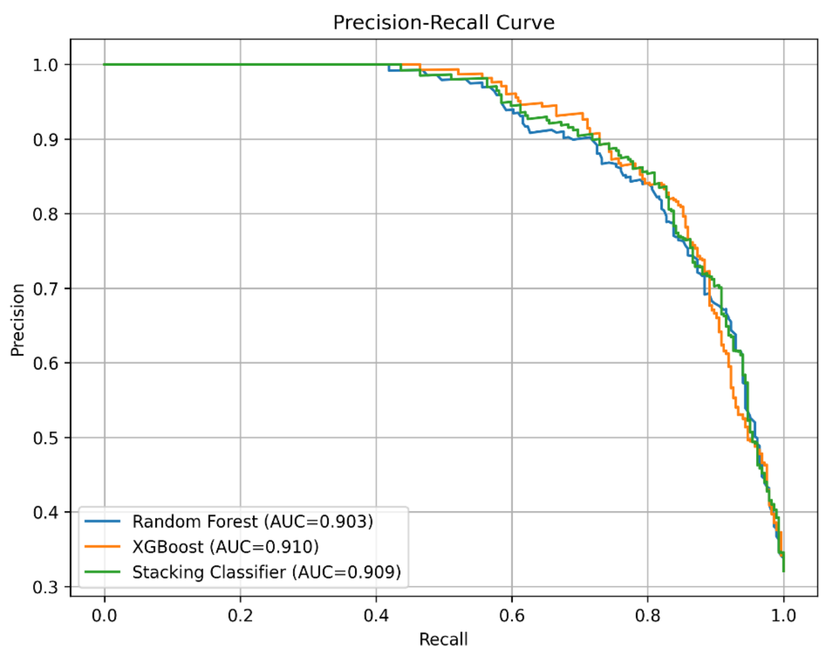

Figure 4. Precision–Recall Curve Analysis for Student Attrition Prediction Models

Recall-Threshold Curve shows that the recall value decreases as the threshold value increases, which can be explained by the increase in confidence required to predict student attrition. A threshold of 0.20 was selected to maximize the number of at-risk students being correctly identified.

From the analysis, Random Forest yields the highest recall, while the Stacking Classifier provides the best balance between the recall and precision at this operating threshold, yielding the best F1-score and the second-best ROC-AUC value. It shows the potential of the Stacking classifier to be a robust decision support tool. Conclusion In this paper, we explore the feasibility of using a stacking classifier for predicting students at-risk of dropping out. The Stacking Classifier based on Random Forest, XGBoost, and Logistic Regression significantly improves the predictive performance over the individual base learners and demonstrates its ability to achieve better prediction reliability and act as a more useful decision support system.

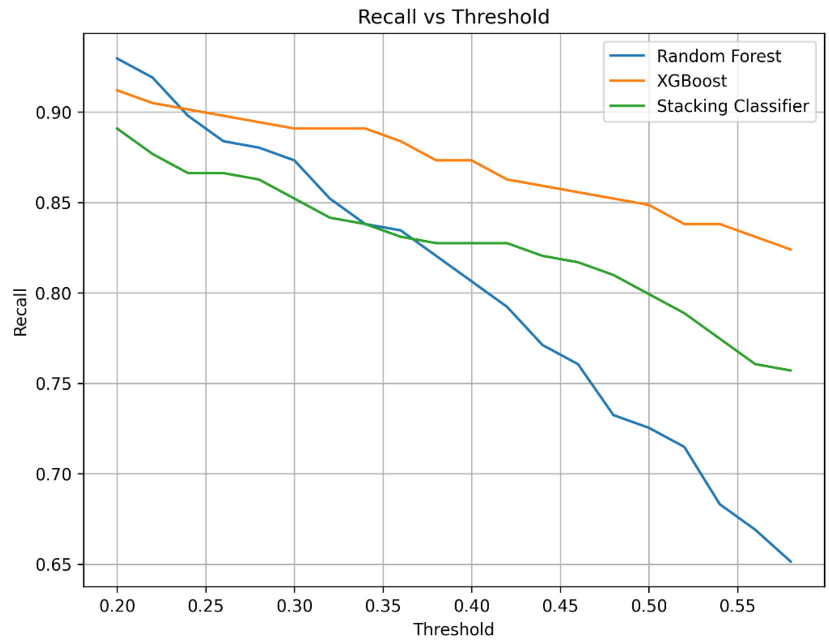

Figure 5. Recall Vs Threshold (Recall as a Function of Classification Threshold for the Evaluated Models)

Table 5. Evaluation Matrix

Model	Accuracy	Precision	Recall	F1-score	PR-AUC	ROC-AUC
Random Forest	0.802	0.630	0.930	0.751	0.903	0.933
XGBoost	0.791	0.618	0.912	0.737	0.910	0.934
Proposed Stacking Classifier	0.852	0.717	0.891	0.794	0.909	0.936

4.3 Risk-Level Prediction and Early Warning System

In addition to the binary classification, the proposed pipeline also provided probabilities of dropout, which were mapped into a categorization of institutional risk levels:

Table 6. Distribution of Students Across Predicted Risk Categories

Risk Category	Number of Students (n)
Low Risk	339
Medium Risk	275
High Risk	271
Total	885

The stratification illustrated the utility of the framework as a part of an early warning system for academic intervention administrators to identify high-risk learners early enough to provide targeted assistance and prevent them from dropping out.

4.4 SHAP & LIME Explainability Analysis

Both the SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) models were used for explaining the predicted drop-out students. SHAP gave a global explanation and found the biggest contributing factor to lower the probability of dropout is academic performance, especially semester 2 curriculum unit approval and payment of tuition fee. LIME provided a local explanation and shown how certain values for a student could lead to predicting a dropout or graduate. Both of these have proven that the model could be relied on to be an explainable and implementable AI to provide a higher education early warning system

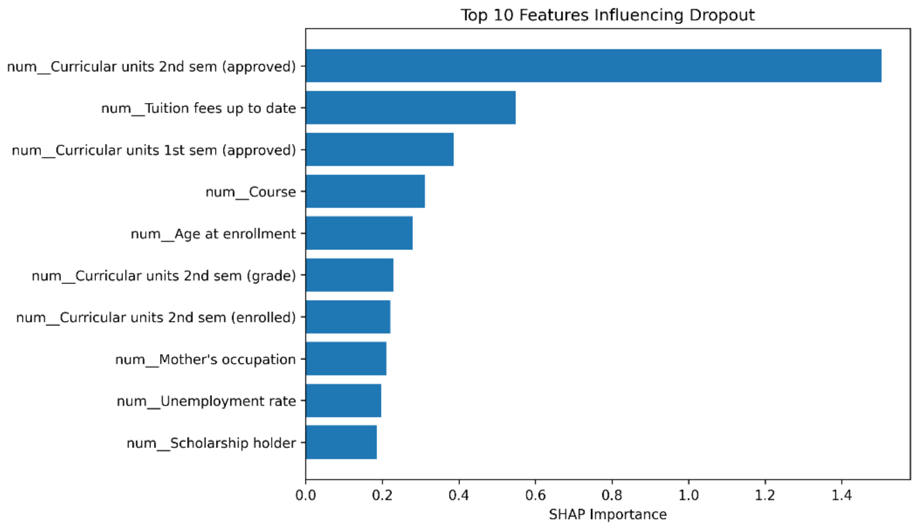

Figure 6. SHAP Summary Plot Illustrating the Most Influential Features for Student Attraction Prediction

Figure 7, shows the LIME explanation of an individual student attrition prediction, provided by our proposed stacking model. We found that the performance features have the largest impact on the model output. The Second-Semester Approved Units and First-Semester Approved Units contribute significantly, and both negatively, suggesting that students with a good academic performance result in a much lower probability prediction of dropout.

Likewise, the fact that the student is non-debtor and his/her young age at the moment of enrolment also lowers the predicted dropout risk, whereas the nationality and academic program both contribute to a smaller extent to the positive side of the predicted outcome.

Generally, the explanation reflects the model decision of the performance of a student, with demographic attributes playing a minor role. The provision of such local explanation helps improve the interpretability and understanding of the decision of the model for a specific student, assisting in the formulation of intervention strategies.

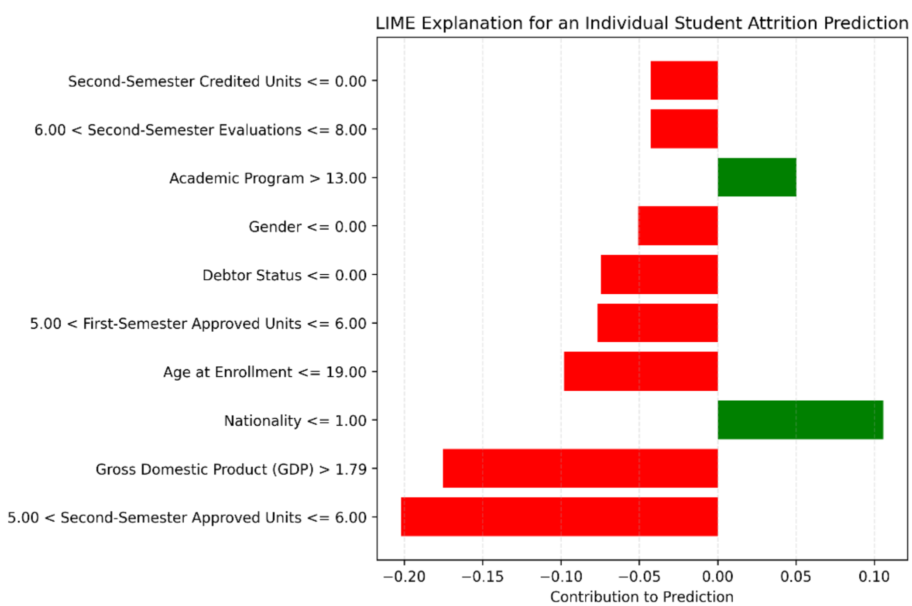

Figure. 7. LIME Summary Plot

5. CONCLUSION

A stacking ensemble learning framework based on explainable artificial intelligence (XAI) for the early prediction of student attrition in higher education was proposed in this work. This framework combines a Random Forest and XGBoost as base learners and a Logistic Regression as a meta-learner, whereas cost-sensitive learning strategy was performed using the parameter scale_pos_weight for class imbalance. Furthermore, we set a classification threshold of 0.20 for improved early prediction of student dropout.

We found through experimental verification that the proposed Stacking Classifier obtained better performance results compared with individual base learners, present-ing a high predictive performance: overall accuracy of 85.2%, precision 71.7%, re-call 89.1%, F1-score 79.4% and ROC-AUC 93.6%.

These results highlight that an ensemble of multiple models in a stacking structure results in better generalization capacity while retaining high student attrition risk identification ability. For model transparency and to gain trust in predictive deci-sion-making process, SHAP and LIME were used to obtain both global and local explanation of the predictions. Our explanations confirm that student performance (approvals), in addition to student financial conditions (debtor status) are among the variables that explain the majority of predictions about student attrition. Based on the predictive probability results, students were categorized as Low, Medium, and High-Risk students by applying a percentile-based risk stratification approach, to enable data-driven prioritisation of interventions.

It is recommended that higher education institutions implement routine monitoring of Low-Risk students, preventive mentoring to Medium-Risk students and targeted academic assistance for High-Risk students.

To summarize, this study demonstrated that by integrating ensemble learning with cost-sensitive optimization, explainable AI, and probability-based risk stratification, our framework offers a comprehensive, transparent and effective strategy for student attrition prediction in higher education institutions that allows for timely decision-making for at-risk students. In the future, this study could be further extended by integrating institutional learning analytics data in real-time, deep learning temporal architectures, multimodal education data, and student historical performances over time to improve the performance of student attrition prediction and robustness across higher education institutions.

Acknowledgements The author sincerely thanks Dr. Rupal Parekh, Atmiya University, Rajkot, Gujarat, India, for her valuable guidance, insightful suggestions, and continuous encouragement throughout this research. The author also expresses sincere gratitude to Thakur Institute of Management Studies, Career Development and Research (TIMSCDR), Mumbai, for providing a supportive academic environment and institutional encouragement while carrying out this research.

References

1. Abhinav P, Ashmitha P, Chakravarthy VS, Makka S (2025) Prediction to intervention: A teacher-friendly early warning system for student dropout. In: Proceedings of the Educational Data Mining Conference, pp 1–10
2. Itani M, Itani MN, Kechli W, Rizk CR (2025) Feature clustering for enhanced student dropout prediction. In: Proceedings of the International Conference on Learning Analytics, pp 1–12
3. Hutagaol N, Suharjito S (2019) Predictive modelling of student dropout using ensemble classifier method in higher education. J Educ Data Min 11(2):45–56
4. P, LR S, V K (2024) Forecasting student attrition using machine learning. In: Proceedings of the International Conference on Artificial Intelligence Applications, pp 1–8
5. Shrirao V, et al (2025) Prediction of student dropout analysis for education using machine learning. In: Proceedings of the Smart Education Systems Conference, pp 1–9
6. Berka P, Marek L (2026) On stability and robustness of students' dropout prediction. J Educ Anal 5(1):21–34
7. Sivakarthi G, Abikannan PR, B T (2024) Predicting student dropout using educational data with temporal convolutional network. In: Proceedings of the International Conference on Machine Learning Applications, pp 1–10
8. Pérez B, Castellanos C, Correal D (2018) Applying data mining techniques to predict student dropout: A case study. Expert Syst Appl 95:128–139
9. Nagy M, Molontay R (2023) Interpretable dropout prediction: Towards XAI-based personalised intervention. IEEE Access 11:55000–55012

10. Prasanth A, Alqahtani H (2023) Predictive modeling of student behavior for early dropout detection using machine learning techniques. Appl Sci 13(4):1–15
11. Sakshi, Sharma C, Kukreja V, Kaur D (2021) Cause analysis of students' dropout rate using PSPP. In: Proceedings of the International Conference on Data Science Applications, pp 1–7
12. Masood S, Gogoi M, Begum S (2024) Optimised SMOTE-based imbalanced learning for student dropout prediction. J Intell Syst 33(2):101–115
13. Wan Yaacob WF, et al (2020) Predicting student drop-out in higher institution using data mining techniques. Int J Adv Comput Sci 11(3):90–99
14. Zerkouk M, Mihoubi M, Chikhaoui B, Wang S (2024) A machine learning-based model for student dropout prediction in online training. Comput Educ Artif Intell 6:100–112
15. Akter T, et al (2024) Dropout prediction of university students in Bangladesh using ma-chine learning. Educ Inf Technol 29(2):2501–2518
16. G V, Eliyas S, M S (2024) Detecting and predicting learner's dropout using KNN algorithm. In: Proceedings of the International Conference on Intelligent Computing Systems, pp 1–6
17. Krüger JGC, Britto A, Barddal JP (2023) An explainable machine learning approach for student dropout prediction. Knowl Based Syst 265:110–120
18. Masrom S, et al (2024) Machine learning prediction for academic misconduct prediction: An analysis of binary classification metrics. IEEE Access 12:34000–34012
19. Mgonja T (2024) Using interpretable machine learning approaches to predict student completion of remedial mathematics. Educ Sci 14(1):1–14
20. Renooij S, Verheij B (2021) Rationale discovery and explainable AI. Artif Intell Rev 54(5):3453–3472
21. Sujon KM, et al (2025) RXK-VEM: A probabilistically calibrated hybrid ensemble model for predicting student dropout in higher education. Expert Syst 42(3):e13123
22. Bassetti E, et al (2022) ISIDE: Proactively assist university students at risk of dropout. In: Proceedings of the International Conference on Learning Analytics and Knowledge (LAK), pp 200–209
23. Rzepka N, et al (2022) Fairness of in-session dropout prediction. In: Proceedings of the ACM Learning Analytics Conference, pp 150–158
24. Cohausz L (2022) Towards real interpretability of student success prediction combining XAI and social science. Educ Inf Technol 27(6):7565–7580
25. Fan T (2026) Enhancing student dropout prediction using deep residual network with optimised feature selection. Neural Comput Appl
26. Mi C (2019) Student performance early warning based on data mining. J Educ Technol Syst 48(2):190–201
27. Revathy SM, Kamalakkannan P, Kavitha R (2022) Machine learning-based prediction of dropout students using SMOTE. Procedia Comput Sci 198:511–518
28. Asha P, Vandana E, Bhavana E, Shankar K (2020) Predicting university dropout through data analysis. In: Proceedings of the International Conference on Data Analytics, pp 1–8
29. Patel D, et al (2024) Empowering student dropout analysis with machine learning and holistic investigation. IEEE Access 12:78000–78015
30. Pacheco ASV, Tete MF, Monsueto SE (2024) Actions to combat student dropout in higher education. High Educ Policy 37(1):75–90
31. Arslan E, et al (2026) Ex-ADA: A SHAP-based explainable AdaBoost framework for predicting at-risk students. Expert Syst Appl
32. Padmasiri P, Kasthuriarachchi S (2024) Interpretable prediction of student dropout using explainable AI models. Appl Artif Intell 38(4):400–415
33. Ashraf A, Anwer S, Khan MG (2018) A comparative study of predicting student performance using data mining techniques. Int J Comput Appl 182(20):15–22
34. Balachandar S, et al (2024) Predicting student dropout rates in MOOCs using an atten-tion-based GCNN model. IEEE Trans Learn Technol
35. Umamaheswari P, et al (2023) Student success prediction using modified SVM-based machine learning approach. Soft Comput 27(18):13211–13224
36. Baumann J, Kramer O (2024) Towards explainable evolution strategies with large language models. In: Proceedings of the International Conference on Artificial Intelligence Research, pp 1–9

37. Kim MG, Jung J (2023) Spectrogram-based explainable ECG analysis for user recognition. Biomed Signal Process Control 86:104–112

38. Marina B, Senthilrajan A (2024) A fuzzy rule and LSOWL–CNN forecasting model for student dropout prediction. Expert Syst Appl 237:120–132

39. Subhiksha SR (2026) A machine learning and explainable artificial intelligence approach to student dropout prediction using multidimensional educational data. Cureus J Comput Sci. https://doi.org/10.7759/s44389-026-00036-8