

Explainable Machine Learning for Student Performance Prediction in Higher Education

Shilpa K

Research Scholar and Associate Professor
Institute of Computer Science and Information Science
Srinivas University, Mangalore
Nrupathunga University, NT road Bengaluru
shilpaktr@gmail.com

Abstract. In higher education the early identification of students who are at risk of academic underperformance is crucial to deliver timely and targeted academic support. In this study, an Adaptive Multi-Modal Learning Analytics Framework (AMLAF) is proposed that combines academic, behavioural, institutional, learning preference and perceptual variables for predicting student academic risk. Explainability analysis with SHAP and LIME revealed that previous academic performance, study behavior, mentoring, institutional support and engagement of students were relevant factors that influence the model predictions. Results showed that the multi modal features engineered using these features were used to train Random Forest, XGBoost and LSTM models, with the latter model proving to be the most accurate. XGBoost with test accuracy of 99.47%, AUC-ROC of 0.9992 and 99.59% mean cross validation accuracy was the best performing model that was observed. The model got 7 false negative predictions and 1 false positive prediction in the test set, and had a recall of 98% and an F1 score of 99% for the at-risk class. The simulated XGBoost–LSTM ensemble obtained an accuracy of 96.75%, which is the same as XGBoost to show no performance improvement over XGBoost alone. LSTM obtained an accuracy of 81.06%, while the simulated XGBoost–LSTM ensemble obtained an accuracy of 96.75%, showing no improvement over XGBoost alone. The results show the first application to supporting early identification of academic risk and evidence based student support planning through multi modal learning analytics, of which explainable is one.

keywords: Academic Risk Prediction; Learning Analytics; XGBoost; Explainable Artificial Intelligence; Student Performance; Higher Education

I. Introduction

Large quantities of data are collected in higher education institutions in digital learning environments, assessment systems, learning management systems and student information systems, concerning academic, demographic, behavioral, and learning data. There is a growing amount of data, especially in the area of learning outcomes, which are generated in higher education institutions in digital learning environments, assessment systems, learning management systems and student information systems, concerning academic, demographic, behavioral, and learning data. With this data, the opportunities for using artificial intelligence (AI) and machine learning (ML) to gain insight into students' learning patterns and predict academic outcomes have arisen. Student performance prediction has therefore emerged as an integral part of the field of educational data mining with the potential to help with academic monitoring, intervention, and personalization, or evidence based institutional decision making. Systematic studies have shown the growth in the use of ML methods when it comes to predicting students' performance, and the algorithms, student related variables, and evaluation methods employed in various educational settings [3, 4, 7, 15].

Traditional methods of measuring student achievement typically rely on grades, classwork, and attendance at school, as well as on the grades of past classes and on teacher evaluations. While these measures are still significant, they are not always useful for determining early on those students who have a high probability of experiencing academic problems. Another data-driven method is machine learning, which can detect non-linear patterns between a number of academic and behavioral factors and be able to predict future performance from these patterns. Early research has demonstrated that ML can be applied to predict student performance and that computational models can help in the assessment of academic performance [6]. A number of comparative studies have been conducted subsequently to assess various ML algorithms, and found that their prediction performance can vary depending on the dataset, features and modelling techniques [11]. In a similar fashion, comparative empirical studies have investigated the different ML models employed for student performance prediction, further highlighting the need for carefully choosing the appropriate model and carefully examining models on a systematic level [8] [10].

Advanced computational procedures have further broadened the realm of pupil predicting. The sophistication of computational methods have also expanded the range of pupil predicting. Recently, ensemble learning, deep learning and hybrid structures have been studied to learn the nonlinear and complex relationship among the educational data. The field of educational prediction problems has been explored recently by the creation of hybrid deep learning models, showing the use of high end neural architectures for student academic performance prediction [1]. Other studies have focused on the use of ML for prediction in university and shown its applications in analyzing academic performance in a university context [16]. The results suggest that the ML models can offer potentially valuable predictive information before implementing major academic tests. Specifically, the application of ML based early prediction can be used to help with formative assessment, and provide educational institutions with an opportunity to intervene early with academic support for students who might need it.

But the achievement of predictive accuracy is not all that is required to effectively use ML in higher education. While many of these advanced models of ML are complex or opaque, the “how” of what they predict might be hard for educators, administrators, and students to grasp. This limitation is especially relevant for the use of predictive outputs to guide academic interventions and/or decisions on student support. Stakeholders should not be given a classification or probability without an explanation of why the student is considered a risk of poor academic performance or is not a risk of poor academic performance if a model predicts a student is at risk of poor academic performance. Explainable artificial intelligence (XAI) aims at this by giving information that is interpretable regarding the contribution and influence that input variables have on the prediction of the model. In the recent systematic review of explainable student performance prediction models, it was emphasized that interpretability in education prediction is becoming more and more relevant and there is a demand for transparency of the prediction system [3].

Incorporating explainability into ML has the potential to enhance the applicability of predictive outcomes in educational settings. For instance, explainable models can show which of the variables that are typically used to predict academic achievement such as previous academic achievement, attendance, assessment performance, learning engagement, and more influences the actual outcome substantially. This data can support teachers in understanding model results and can enable teachers to tailor academic supports even more. The application of explainable AI with machine learning (ML) to the prediction of student performance patterns in practice has been proven effective in early identification of these patterns [12]. Recent studies have specifically explored the use of ML to predict academic performance and provide explanations for it, in the context of education, to aid decision making in schools [2]. In the same vein, the focus on interpretability and education decision making has stressed the importance of insights gained into factors that influence academic predictions more than purely the performance based prediction measures [9].

Explainability also plays a key role in transparency, fairness, and the acceptance of the use of AI in education decision making. The potential usefulness of a highly accurate prediction model may still be limited if the model's predictions are not easily explained or understood by the user of the model. It may be necessary for educational stakeholders to consider the meaning of predictions that are based on some sort of academic indicator and the meaning of explanations that follow these predictions within the context of that particular institution. The experimental study on explainable AI for academic performance prediction investigated how the accuracy and simplicity of the decision tree influenced the students' perception of causability and fairness of the decision tree, showing that the characteristics of an explanatory model can influence users' perception of its decision [13]. This underscores the need for taking interpretability into account in designing AI-based educational systems as well as their predictive capabilities.

The relevance of explainable prediction is further illustrated when ML models are used to predict for certain students and in specific educational settings. The learning contexts and needs, as well as the academic engagement, of various groups of students may differ. Explainable ML can also offer explanations of the factors that impact predicted performance and help to improve the interpretability of the model's predictions. Explainable ML has been used to predict academic

performance of deaf students, for example, in the context of students with unique educational contexts [14] that shows that explainable predictive approaches are relevant to specific student populations. While there have been significant advances in the ability to accurately predict student performance using ML, there are still some research challenges that need to be addressed. Various data sets, learning settings, feature sets, algorithms and metrics have been used in the past, and it is hard to draw conclusions about the relative effectiveness of predictive approaches [4, 7, 15].

1. Research Objectives

The research objectives of this study are,

- To assess the predictive accuracy of machine learning models for predicting students' performance in Higher Education.
- To recognize and describe the important factors that affect student performance predictions, and analyze the value of these factors in educational decision making.

2. Significance of the Study

This research helps the field of higher education analytics by proposing an explainable machine learning approach to accurately and interpretably predicting a student's performance. It allows the identification of influential academic, behavioral and demographic predictors and helps to overcome the opacity of traditional predictive models. The layering of explainability techniques improves model transparency, interpretability and trust by educators and administrators. The results can facilitate evidence based learning and teaching decision making, early identification of academically vulnerable students, use of interventions, individual learning paths, and allocation of resources in institutions in a responsible and transparent way in the field of educational data analytics.

II. Related Work

Lünich and Keller (2024) looked at the effects of decision tree parameters such as model simplicity and predictive precision on perceptions of distributive and informational justice. The results showed that a more simple decision tree had a positive impact on perceptions of fairness, which was mediated by perceived causability. The research emphasizes the need for model transparency and interpretability in the prediction of academic performance using machine learning. Alkan et al. (2025) created a machine learning powered early warning system for student at-risk of academic failure and early intervention. Random Forest algorithm was the best performer when evaluating the algorithms with the cross validation measures. The study highlights the importance and potential of machine learning based early warning systems to support proactively academic interventions. To overcome the drawbacks of traditional academic performance prediction methods, Guevara-Reyes et al. (2025) proposed the use of machine learning models optimized with techniques of XGBoost and Random Forest. While studies that focus on prediction mostly center on the level of prediction, the research included interpretability techniques based on the SHAP model to see which variables contributed to the model's prediction. The results show the importance of explaining and predictive modelling approaches in education.

Raji et al. (2024) created a machine learning based model to forecast the academic risk of deaf and hard of hearing (DHH) students using explainable machine learning (XAI). The proposed method combined the machine learning model with Explainable Artificial Intelligence (XAI) techniques: the Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP). The study illustrates how a combination of predictive modelling and interpretable explanations can lead to socially relevant insights, and help educational stakeholders to create targeted strategies that will enhance the academic outcomes of students with specific educational needs. Heikal and Khalil (2023) comprehensively evaluated all the machine learning methods applied to the Student Performance Dataset in the prediction of student performance. A comparative analysis of the different machine learning models that were applied provided useful insights concerning the effectiveness of the various models for the student performance prediction tasks. Researchers Ahmed et al. (2025) studied the academic performance prediction using machine learning to overcome challenges of student evaluation, academic performance analysis and decision making for educational purposes based on evidences. Academic outcomes were predicted and factors associated with student performance were identified using the models.

Soyoye et al. (2023) used machine learning algorithms to select factors that are correlated to university students' academic performance and to aid in the early detection of students who need academic support. Based on the accuracy of the evaluation results, Decision Tree and AdaBoost regression models can be said to have a higher degree of accuracy than the other models. The study illustrates how machine learning can be used in institutional academic records to identify students that can benefit from targeted help. Wishing to explore the strengths and weaknesses of the different machine learning techniques, Chen and Zhai (2023) carried out a comparative study of machine learning approaches with three different kinds of task oriented educational datasets. The results of the experiments showed that the Random Forest algorithm had good generalization properties in the datasets chosen. The importance of comparative evaluation of models for different educational prediction tasks and datasets is highlighted. Albreiki et al. (2021) performed a systematic literature review of the studies on student dropout and academic risk prediction that have been published from 2009 to 2021. This review focused on the use of different machine learning methods to detect students at risk for poor academic performance and/or attrition. The results revealed that data sources in research are mostly from University databases and online learning platforms.

The authors of the systematic review by Balaji et al. (2021) systematically reviewed machine learning models for predicting academic performance of students. The review confirmed that machine learning methods could be used to make good prediction of student performance by using certain academic and behavioral attributes. The study also pointed out the education data mining is not well developed and further study is needed to have more convincing, generalizable and comparable results. This systematic review by Alamri and Alharbi (2021) covers explainable models for student performance prediction that have been published from 2015 to 2020. The results showed that there was a need for additional research that would consider both the accuracy of the predictions and the explainability of the model with systematic and measurable criteria. The researchers Jang et al. (2022) suggested a feasible prediction model and method that combines machine learning and Explainable Artificial Intelligence (XAI) techniques. The potential features that can be used in the prediction model were selected at the focus group discussions and

then further determined by correlation analysis. The results highlight the need to take into account the views of all stakeholders in choosing meaningful predictive features and in creating interpretable student performance prediction systems.

Adefemi et al. (2025) suggested a hybrid deep learning model that combines a Convolutional Neural Network (CNN) and a Bidirectional Gated Recurrent Unit (BiGRU) for predicting student academic performance. The architecture proposed is targeted to enhance the predictive performance by capturing the complex patterns in the student related data. The study also examined the critical issues of preprocessing such as missing values, class imbalance and feature selection. Ismail et al. (2021) compared some of the popular machine learning classification models for determining students that are likely to have learning difficulties. This study intended to add to the early identification process and the use of the appropriate precautionary measures. The results revealed that the performance of the models was dependent on the characteristics of the dataset, highlighting that model selection should be done on an empirical basis, instead of generalizing that one model is best suited for all datasets. Oppong (2023) carried out a systematic review of literature on machine learning techniques adopted for student performance prediction with a special focus on developing significant attributes in the student data sets. The results provide evidence of the growing application of sophisticated machine learning techniques for the detection of patterns related to learning outcomes.

Altabrawee et al. (2019) studied the impact of technology related behavior on students' academic achievement based on machine learning techniques. Special emphasis was placed on the students' use of the internet as a learning tool, as well as how much time they spent on social networking activities. The paper provides an example of the usefulness of behavioral and technology use variables to build models that predict student performance. Using three datasets of different tasks, Chen and Zhai (2023) also conducted a comparative study of the performances of machine learning algorithms in different educational prediction tasks. The evaluation involved seven machine learning methods optimized for each one of the parameters in both binary and multiclass classification problems. The results demonstrated that Random Forest had good generalization performance on the selected datasets, indicating ensemble based methods may have good performance under different educational prediction scenarios.

1. Research Gap

The literature study conducted showed significant advances in the use of machine learning in student performance prediction, early alert systems and academic decision support. There are however some restrictions. Numerous studies focus mainly on predictive accuracy, and a number of studies that include some aspects of interpretability and/or educational usability do not do so in an integrated framework. While there has been an increasing number of XAI techniques applied, there is limited research as yet on how to measure the quality of the explanations and its utility for educational stakeholders. Moreover, there are various dataset, feature, educational and assessment differences between models, making direct comparisons difficult. Thus, there is a need for an integrated explainable machine learning framework for higher education that is used to predict accurate student performances and provide explanations for individual prediction of student performances.

III. Research Design and Methodology

1. Research Design

To build and empirically test an explainable machine learning framework for student performance prediction in higher education, a pragmatic, quantitative predictive analytics design was coupled with those of Design Science Research (DSR). The methodological framework aimed to fulfill two complementary needs: predictive accuracy and interpretability. Therefore, a combined analytical pipeline that includes conventional machine learning, deep learning, explainable artificial intelligence (XAI), fairness evaluation, and cross institutional validation, was included.

The proposed Adaptive Multi-Modal Learning Analytics Framework (AMLAF) comprised of three main stages: (i) multi-institutional data infrastructure, (ii) predictive modelling and evaluation, and (iii) explainability and fairness assessment. The research process in table 1 was carried out systematically from data acquisition and integration, to data preprocessing, feature engineering, model development, XAI based interpretation, and thorough validation.

Table 1. Research Design Structure

Component	Methodological Specification
Research paradigm	Pragmatism
Research design	Quantitative predictive modelling and design science research
Unit of analysis	Individual higher education student
Study setting	Government affiliated higher education institutions in Karnataka
Institutions	20
Initial sample	10,065 students
Final modelling sample	10,035 students
Prediction task	Binary academic-risk classification
Prediction point	End of Semester 2
Outcome	Semester 3 SGPA
At-risk criterion	SGPA < 6.5
Predictive models	Random Forest, XGBoost, LSTM
Explainability methods	SHAP and LIME
Validation	10-fold stratified CV and LOIO-CV
Additional assessment	Fairness, actionability, and statistical significance

2. Data Collection and Variables

Data in table 2 are primary data, which were obtained using an online structured questionnaire in November 2024. The instrument was comprised of 25 items and took about 10 minutes to complete. Information about demographic, academic, behavioral and institutional support variables were obtained through the questionnaire. The academic predictors were 10th grade percentages, 12th grade percentages, and first semester and second semester of the student's Grade Point Average (SGPA). Behavioral variables consisted of weekly study hours and learning style preference, stress level, goal setting behavior and use of external learning materials. The institutional

variables were the mentoring availability, awareness of student support services, academic monitoring satisfaction, and perceived institutional support.

The academic risk indicator was operationalized as a binary variable (academic risk/no academic risk) using the third semester SGPA:

$$Y_i = \begin{cases} 1, & SGPA_{Sem3} < 6.5 \\ 0, & SGPA_{Sem3} \geq 6.5 \end{cases} \quad (1)$$

Table 2. Operationalization of Study Variables

Variable category	Principal variables	Measurement
Academic history	10th %, 12th %, SGPA1, SGPA2	Continuous
Academic trajectory	Pre-university average, consistent performer	Derived
Study behaviour	Study hours, low/high study hours	Continuous/binary
Well-being	Stress, high/low stress	1–10/binary
Institutional support	Mentoring, academic monitoring, support awareness	Binary/ordinal
Learning behaviour	Learning style, external study-material use	Categorical/ordinal
Demographics	Gender, degree level, institution	Categorical
Outcome	Third semester SGPA	Continuous
Prediction target	Academic risk status	Binary

3. Data Preprocessing and Feature Engineering

Data preprocessing involved tasks such as data quality control, handling missing data, fixing data types, data range checking, outlier detection and removal, categorical encoding and prevention of leakage. Table 3 shows only 56 of 271,755 data cells were missing (0.02%). Continuous variables were imputed using median and categorical variables were imputed using mode method. All non-numeric values given in the SGPA fields were coded as missing values prior to imputation. To keep the data within a realistic range, the data on the SGPA was constrained by domain constraints to the range 0–10. The procedure of interquartile range was used for assessment of outlier. Data values for study hours that were outside of reasonable bounds were replaced with a value rather than removed in order to retain observations. Interpretable variables for academic trajectory, behavioral patterns, stress, institutional support and interactions were created through feature engineering. Thirty predictive features were chosen for modelling.

Table 3. Feature Engineering and Encoding Strategy

Feature group	Examples	Processing
Academic	SGPA1, SGPA2, pre-university average	Continuous/derived
Academic consistency	Consistent performer	Binary
Study behaviour	Study hours numeric, low/high study hours	Continuous/binary
Stress	High stress, low stress	Binary
Institutional support	Mentoring indicators	Binary
Learning style	Visual, Reading, Writing, Practical	One-hot
Demographics	Gender, degree level	Binary
Interaction	Study × stress, SGPA × stress	Continuous
Institution	Institution identifier	Encoded categorical

4. Predictive Model Development

Table 4 present the cleaned dataset was split into subsets of 70% training, 15% validation and 15% testing, and stratified to retain the class distribution of 28:72. To achieve the reproducibility, a fixed seed number '42' was used for the random number generators. Only information available at the end of Semester 2 were used for all the predictors. The following three different models were tested: Random Forest, XGBoost, and Long Short-Term Memory (LSTM). Ensemble based machine learning (Random Forest and XGBoost) was appropriate for structured educational data, while deep learning (LSTM) was appropriate for sequential academic information.

Table 4. Predictive Model Configuration

Model	Principal configuration	Purpose
Random Forest	100 trees; min samples split = 5; min samples leaf = 2; balanced class weights	Ensemble tabular classification
XGBoost	100 estimators; depth = 6; learning rate = 0.1; subsample = 0.8; feature sampling = 0.8; scale_pos_weight = 2.57	Gradient boosted classification
LSTM	2 LSTM layers $\times$ 128 units; dropout = 0.3; Adam, LR = 0.001	Temporal sequence modelling

5. Explainable Artificial Intelligence

The explainability was added by incorporating it using SHAP and LIME to get both global and local interpretations of predictive output, as presented in Table 5. TreeSHAP was used to create SHAP values for the tree based models. Overall feature importance was quantified with mean absolute SHAP values and individual feature values were used to quantify feature level contributions to specific student predictions. As a local explanation method, LIME was used, which is a model agnostic approach. This approach yielded noisy observations for individual students and learned an understandable local surrogate model to mimic the behavior of the underlying classifier. The 10 most influential features were selected for making individual predictions.

Table 5. Explainability Framework

XAI method	Analytical level	Primary purpose
SHAP	Global and local	Feature attribution and contribution magnitude
SHAP summary	Global	Overall predictor importance
SHAP dependence	Global	Non-linear and interaction effects
SHAP waterfall	Local	Individual prediction explanation
LIME	Local	Instance specific surrogate explanation
Actionability analysis	Decision level	Identification of modifiable predictors

6. Model Evaluation and Validation

The accuracy, precision, recall, F1-score, AUC-ROC and confusion matrices were used to evaluate the model performance as shown in Table 6. Failure to recognize an academically vulnerable student could lead to a lack of opportunity for timely intervention, so recall for the at-

risk class was seen as a special diagnostic measure. To check the internal stability, Cross Validation (CV) with ten folds was used for training set. Furthermore, Leave-One-Institution-Out Cross-Validation (LOIO-CV) was performed over all 20 institutions to evaluate the generalizability of results to external institutions. In every round, one institution was left aside for testing and the others institutions were used for modeling development.

Table 6. Model Validation Protocol

Validation component	Procedure	Purpose
Hold-out testing	70:15:15 train/validation/test	Unbiased final evaluation
10-fold CV	Stratified training-set cross-validation	Internal stability
LOIO-CV	One institution held out per iteration	Cross institutional generalizability
Confusion matrix	TP, TN, FP, FN	Error analysis
AUC-ROC	Threshold independent discrimination	Class separability
Fairness analysis	Demographic and institutional groups	Bias assessment

IV. Results and Analysis

1. Model Performance

A sample of the performance of the models evaluated is presented here in Table 7. Test Accuracy for both the baseline models was greater than 99% making the performance of the models highly predictable. XGBoost outperformed Random Forest with test accuracy of 99.47% and test AUC of 0.9992 while Random Forest came up with 99.14% and 0.9991 respectively. The difference between 10-fold cross validation mean for XGBoost (99.59%) and Random Forest (99.22%) suggest that XGBoost has a stable prediction performance across the validation folds. Both models experienced only a few percentage points of train–test performance degradation, with the difference between the two less than 1%. XGBoost was thus the top performer overall and detailed analysis was conducted using this method.

Table 7. Baseline Model Performance

Model	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Validation AUC	Test AUC	CV Mean (%)
Random Forest	99.73	99.47	99.14	0.9996	0.9991	99.22
XGBoost	100.00	99.67	99.47	0.9993	0.9992	99.59

2. XGBoost Classification Performance

The model, XGBoost, was tested using 1510 test observations with 1087 not at risk and 423 at risk students. The result of classification is displayed in Table 8. The confusion matrix analysis results showed that there were seven false negatives and one false positive, making a total of eight classification errors. Hence, 1502 out of 1510 observations were predicted correctly. The recall was about 98.34% accurate as 416 out of 423 students were correctly identified among the at-risk group. For 'not-at-risk' students, there were 1,086 correct classifications of 'not-at-risk' students out of 1,087, with a specificity of ~99.91%. The results of this study show that the main

misclassification error was that of seven at-risk students being classified as not at-risk. The small number of false positives (fewer than 10) suggests that there were also few unnecessary risk alerts in the test set. Figure 1 shows XGBoost Confusion Matrix (Test Set).

Table 8. XGBoost Test-Set Classification Performance

Class	Precision	Recall	F1-score	Support
Not At-Risk	0.99	1.00	1.00	1,087
At-Risk	1.00	0.98	0.99	423
Weighted Average	1.00	0.99	0.99	1,510

Confusion Matrix Analysis

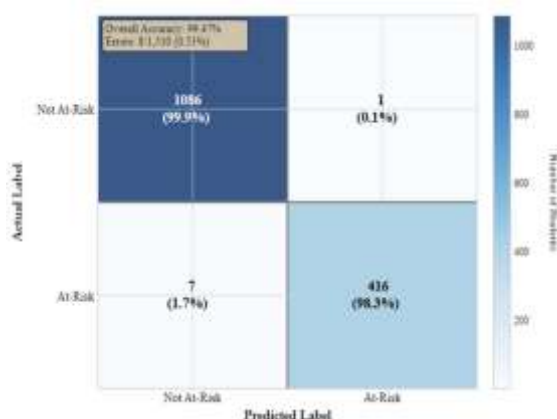


Figure 1: XGBoost Confusion Matrix (Test Set)

Feature Importance and Explainable AI Analysis

The top predictors of the gain-based XGBoost analysis were institutional, academic, behavioral, learning preference, support, and perceptual variables. Table 9 lists the top 10 of the features. In addition to the traditional academic variables, the institutional and behavioral variables were also important in the predictive structure. Instruments of support information such as mentoring related variables were also presented in the task classification repeatedly among the high importance predictors, highlighting the importance of such information in the classification task.

Table 9. Top XGBoost Features by Gain Based Importance

Rank	Feature	Importance (%)	Category	Actionability
1	Industrial visits	6.23	Institutional	Yes
2	Visual learning	5.06	Learning Style	Partial
3	Pre-university average	4.12	Academic	No
4	High study hours	4.08	Behavioral	Yes
5	1st Semester SGPA	4.07	Academic	No
6	Mentoring encoded	3.78	Support	Yes
7	Has mentoring	3.76	Support	Yes
8	Study hours	3.71	Behavioral	Yes

9	2nd Semester SGPA	3.68	Academic	No
10	12th percentage	3.66	Academic	No

As shown in table 10, 33.8% of the reported gain based importance was observed for the category level behavioral and institutional support variables. These variables are factors that can be potentially addressed through academic support, mentoring, study planning and institutional interventions.

Table 10. Feature Importance by Category

Category	Total Importance (%)	Actionability
Academic Performance	28.5	Low
Behavioral Factors	15.2	High
Institutional Support	18.6	High
Learning Preferences	14.6	Medium
Perceptual/Attitudinal	11.4	Medium
Demographics	11.7	Low

LSTM Temporal Model Results

The test accuracy of LSTM model is 81.06% with the AUC-ROC is 0.8847 as shown in Table 11. Early stopping was implemented and 28 epochs were needed for training. The model had some 229,000 parameters.

Table 11. LSTM Model Performance

Metric	Result
Test accuracy	81.06%
Test AUC-ROC	0.8847
Best validation accuracy	83.5%
Training epochs	28
Parameters	229,000
CPU training time	27 minutes
GPU training time	3 minutes

In comparing the performance of the LSTM with the XGBoost, the results were significantly inferior with LSTM, especially in terms of identifying at risk students, as demonstrated in table 12.

Table 12. LSTM Classification Performance

Class	Precision	Recall	F1-score	Support
Not At-Risk	0.81	0.95	0.88	1,087
At-Risk	0.79	0.44	0.57	423
Macro Average	0.80	0.70	0.72	1,510
Weighted Average	0.81	0.81	0.79	1,510

The most significant difference was the significantly reduced at-risk recall of LSTM. Table 13 shows that the model failed to identify the 237 of the 423 at-risk students while XGBoost missed 7 students. The difference makes sense based on the structure of the data: the LSTM took in four academic observations over time, along with a set of context variables, while XGBoost used a larger engineered feature set, including academic, behavioral, institutional and contextual features.

Table 13. XGBoost–LSTM Comparison

Metric	XGBoost	LSTM	Difference
Test accuracy	99.47%	81.06%	18.41 percentage points
AUC-ROC	0.9992	0.8847	0.1145
At-risk precision	1.00	0.79	0.21
At-risk recall	0.98	0.44	0.54
At-risk F1-score	0.99	0.57	0.42
False negatives	7	237	230
False positives	1	55	54

SHAP Analysis

In table 14, SHAP was used to give explanations for the XGBoost predictions globally. Semester performance and prior academic achievement were the top two variables with the largest mean absolute SHAP values, followed by behavioral and institutional variables. The results of the SHAP dependence analysis showed a negative correlation between the increase in first semester SGPA and the academic risk predicted. The reported dependence analysis also revealed an inflection around the SGPA of ~6.5. Study hours were found to be related in a non-linear manner with less study duration predicted to be higher risk and effects were found to gradually reduce at higher study durations.

Table 14. Top 15 Features by Mean Absolute SHAP Value

Rank	Feature	Mean Absolute SHAP
1	1st Semester SGPA	0.7741
2	10th Percentage	0.6686
3	12th Percentage	0.6497
4	2nd Semester SGPA	0.6142
5	Pre-university average	0.5987
6	Study hours	0.2932
7	Quality insights	0.2680
8	Frequency of external-material use	0.2303
9	Mentoring	0.2152
10	Awareness of support	0.2056
11	Goal setting	0.2001
12	Global certification	0.1914
13	High stress	0.1749
14	Satisfaction	0.1651
15	Visual learning	0.1029

As a consequence, the reported SHAP grouping in table 15 suggests that around 31.2% of the contribution of the model is due to actionable behavioral and institutional variables, whereas academic history is responsible for the biggest contribution.

Table 15. SHAP Importance by Actionability

Category	Total SHAP	Percentage of Total	Actionability
Academic History	3.3054	65.0%	Low
Behavioral Factors	0.8985	17.7%	High
Institutional Factors	0.6888	13.5%	High
Demographics	0.1701	3.3%	Low
Learning Styles	0.2901	0.6%	Medium

Local Explainability: SHAP and LIME

The individual SHAP and LIME analyses in table 16 illustrated the interpretation of predictions at student level. The positive factors contributing to the high confidence at-risk SHAP example were a low first semester SGPA, low study hours, lower prior academic performance and not having mentors. In contrast, students who had higher grades in the previous school year had a negative effect on the risk prediction. The LIME explanation for Student #1523 also included low 10th grade performance, low study hours, lack of mentoring, and low first semester SGPA as important local factors. The top ten most important features by SHAP and LIME have 3/10 (30%) shared features. The difference arises from the different interpretations that can be given to the two approaches: SHAP can be used to explain the model's behaviour on a whole dataset, while LIME can be used to give an approximation of the model's behaviour on a single prediction.

Table 16. SHAP–LIME Feature Overlap

Explanation Method	Principal Features Identified
SHAP	1st Semester SGPA, 10th %, 12th %, 2nd Semester SGPA, pre-university average, study hours, quality insights, mentoring
LIME	Frequency of external materials, support awareness, practical learning, mentoring, study hours, quality insights
Common features	1st Semester SGPA, study hours, mentoring

Institution-Level Analysis

Table 17 showed that there were differences in the relative importance of the predictors across institutions in the institution specific SHAP analysis. In 17 of the 20 institutions, the first semester SGPA was one of the top 10 predictors. Other high ranking predictors were institution related (institutional quality variables) as well as other academic-related variables (previous academic performance, mentoring, study hours), depending on the institution. The results at the institution level showed that despite academic progression constantly playing an important role, there were differences in the secondary predictors among institutions. This helps to enable the use of institution specific interpretation in translating the model outputs into intervention strategies. This is a SHAP Feature Importance Heat map of 20 Institutions (Figure 2).

Table 17. Examples of Institution-Specific Feature Importance

Institution	Rank 1	Rank 2	Rank 3
College A	1st Semester SGPA	10th %	Study hours
College B	1st Semester SGPA	Mentoring	12th %
College C	1st Semester SGPA	Quality insights	2nd Semester SGPA
College D	10th %	1st Semester SGPA	High study hours

Heat map across 20 Institutions

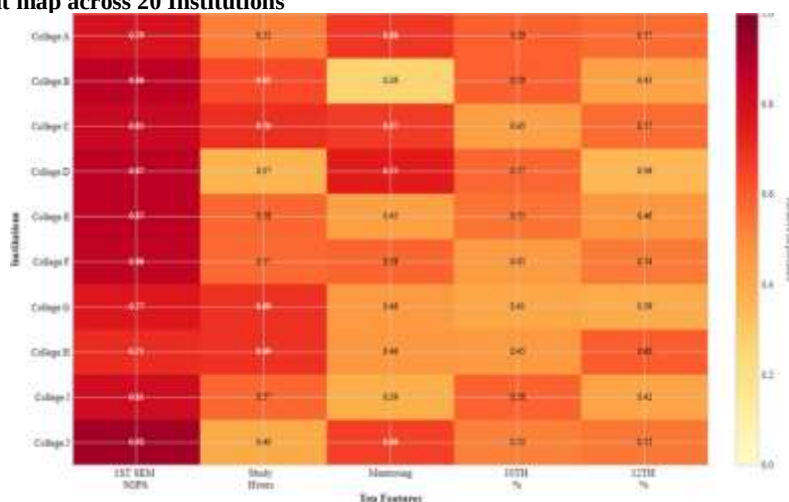


Figure 2. SHAP Feature Importance Heat map across 20 Institutions

Ensemble Analysis

To see if there was an improvement in the observed predictive performance, an ensemble analysis was simulated, the results of which are shown below. An ensemble analysis was simulated to see if there was an improvement in the observed predictive performance, as shown below. In the reported simulation, the combination of XGBoost and LSTM gives an accuracy of 96.75% while the XGBoost gives an accuracy of 99.47%. The simulation was thus not able to show the advantage in terms of accuracy of the prediction with the addition of LSTM under the weighting method used.

Table 18. Ensemble Simulation Results

Scenario	Partner Model	Partner Accuracy	Expected/Observed Ensemble Accuracy
Baseline	XGBoost only	99.47%	99.47%
Scenario 1	Perfect partner	100.00%	100.00%
Scenario 2	Strong partner	95.00%	99.30–99.50%
Scenario 3	LSTM	81.06%	96.75%*

*Reported simulated result using a 90% XGBoost and 10% LSTM combination.

V. Findings and Discussion

The results show that the proposed Adaptive Multi-Modal Learning Analytics Framework (AMLAF) resulted in very good predictive performance in identifying students who are at risk of underperforming in learning. The test accuracy and AUC-ROC of XGBoost was 99.47% and 0.9992, respectively, based on the data from 10,065 students from 20 institutions, and the mean test accuracy of Cross validation was $99.59\% \pm 0.16\%$. The model was able to correctly classify nearly every student in the not at risk and at risk categories on the test set with just 8 classification errors (7 false negatives and 1 false positive). The at-risk class had a recall score of 98% and F1-score of 99% which means that the model was able to identify most of the students that needed academic risk attention. This performance implies that the academic, behavioural, institutional, learning preference and perceptual variables that were engineered conveyed a significant amount of predictive information for the classification of academic risk in third semester.

Feature-importance and explainability analyses additionally helped to explain the influence of the factors in model prediction. The most important predictors, according to the gain based XGBoost importance indicated, were institutional engagement, learning preferences, previous academic performance, study behaviour, mentoring and student support variables. First semester SGPA, percentages at 10th and 12th grades, second-semester SGPA, and pre-university performance were other factors identified by the SHAP analysis to predict the students' grades. The results suggest that the academic achievement factor was still significant in predicting risk, and that the other factors measured the behavioural and institutional-support were additional factors that could help guide action. Importantly, these importance values are not causal, but rather they are contributions to model predictions. The fact that a significant amount of predictive information can be found in earlier academic performance also implies that academic risk can be identified prior to significant performance losses in later semesters.

The behavioral/business aspects are especially important in the intervention perspective. The gain based feature importance showed that behavioural factors and institutional support accounted for around 33.8% of the reported gain based feature importance; the same features, along with study hours, stress, goal setting, mentoring, and institutional support indicators proved to be meaningful from the SHAP based analysis. SHAP dependence analysis indicated that study hours had non-linear relationships with predicted risk with lower 1st semester SGPA generally related to higher predicted risk. Evidence at the student level was also supplemented by local explanations using LIME which found that low student first semester SGPA, limited study hours, lack of mentoring and low use of external learning resources were important explanatory factors for a high risk case. However, the similarity between SHAP and LIME was small: they highlighted three variables of the ten, due to their different purposes in analysis SHAP offers a contribution framework which is global or theoretically consistent, while LIME provides a local explanation of individual predictions.

Significant differences were found between the modelling approaches when compared to LSTM. The test accuracy of XGBoost and LSTM was 99.47% and 81.06% while the AUC of XGBoost was 0.9992 against 0.8847 of LSTM. Compared to XGBoost which made 55 false-positive and seven false-negatives, the LSTM made 237 false-positive and 55 false-negative

cases. This difference corresponds to the tabular structure and mainly structured nature of the data and the limited time granularity that the sequential model can handle. The simulated XGBoost–LSTM ensemble model attained 96.75% accuracy with 2.8 ms inference time while XGBoost was able to achieve the accuracy of 99.47% with the inference time of 0.3 ms. In this case, the combination of these two models did not lead to a better predictive performance on this evaluation.

The results of the fairness analysis revealed fairly minimal differences in accuracy based on gender and degree. The accuracy was 99.52% for males and 99.41% for females, with a 1.9% difference between the at-risk false-negative rate for males and females. The accuracy of the institutions ranged from 97.8% to 100.0% with the average accuracy of 99.3%. The results suggest that the performance of predictions is broadly similar in the subgroups assessed and that false negatives are of concern and should be monitored and analysed in subgroups.

Study Limitations

There are some restrictions to be taken into account when reading these results. The data pertain to the 20 institutions involved in the sample, which is drawn from the setting of government colleges in Karnataka, which limits the external validity of the findings to the case of private institutions, other Indian states or other international higher education systems. Second, the high prediction performance of our model, XGBoost, might be due to the high predictive power of the previous academic achievement and the outcome measured by the SGPA in the third semester. Independent external validation is therefore needed to see if similar performance is upheld with other cohorts and institutions. Third, the LSTM model was only provided with the time information made available, which is limited and prevented the modelling of long term student trajectories. Fourth, the subgroup analyses (especially at individual institution level) might be sensitive to the sample size. Lastly, both SHAP and LIME give explanations of the models and do not provide evidence of causality. Future studies should therefore factor in prospective multi-institutional validation and leave-in-institution testing, as well as calibration analysis and more exploration of fairness and causal intervention effects over time across successive cohorts of students.

VI. Conclusion and Future Scope

The study aimed to design and test the Adaptive Multi-Modal Learning Analytics Framework (AMLAF) for early detection of students' academic risk in higher education. XGBoost, with data from 10,065 students from 20 institutions, performed well with 99.47% of test accuracy and 0.9992 AUC-ROC, showing it has a good predictive power of students who are at risk of academic underperformance. SHAP and LIME analyses provided an enhanced view of the levels of interpretability through the identification of the contributions of previous academic achievement, study behaviour, mentoring, institutional support, and related student characteristics. The comparison with LSTM and the simulated ensemble also demonstrated that the performance of the standalone XGBoost model outperformed the other models in the reported evaluation in terms of the predictive accuracy and inference time. The accuracy of the evaluated demographic and institutional groups were generally consistent based on a fairness analysis, but there were some differences in false-negative rates that need to be monitored.

Further studies are needed to confirm AMLAF in a prospective fashion in other institutions, states, educational sectors and student groups to confirm external generalisability. Leave one institution out and temporal validation should be carried out to evaluate robustness to distributional change. Rich learning trajectories, attendance data, assessment data, learning management system (LMS) interactions, and interventions data can be added to future studies. Model calibration, subgroup fairness, and causal analyses should also be explored to determine any differences between predictive associations and those that could be used to inform effective interventions. Last but not least, deployment studies should assess the effect of explanations of predictions on timely academic support decisions and student outcomes in actual schools.

References

1. Adefemi, K. O., Mutanga, M. B., & Jugoo, V. (2025). Hybrid deep learning models for predicting student academic performance. *Mathematical and Computational Applications*, 30(3), 59. <https://doi.org/10.3390/mca30030059>
2. Ahmed, W., Wani, M. A., Plawiak, P., Meshoul, S., Mahmoud, A., & Hammad, M. (2025). Machine learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Scientific Reports*, 15(1), 26879. <https://doi.org/10.1038/s41598-025-12353-4>
3. Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, 9, 33132–33143. <https://doi.org/10.1109/ACCESS.2021.3061368>
4. Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student' performance prediction using machine learning techniques. *Education Sciences*, 11(9), 552. <https://doi.org/10.3390/educsci11090552>
5. Alkan, B. B., Kuzucuk, S., Alkan, N., et al. (2025). Using machine learning to predict student outcomes for early intervention and formative assessment. *Scientific Reports*, 15, 39797. <https://doi.org/10.1038/s41598-025-23409-w>
6. Altabrawee, H., Ali, O., & Qaisar, S. (2019). Predicting students' performance using machine learning techniques. *Journal of University of Babylon for Pure and Applied Sciences*, 27(1), 194–205. <https://doi.org/10.29196/jubpas.v27i1.2108>
7. Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: A systematic review. *Applied Sciences*, 11(21), 10007. <https://doi.org/10.3390/app112110007>
8. Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Education and Information Technologies*, 28(9), 12039–12057. <https://doi.org/10.1007/s10639-023-11672-1>
9. Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F., & Villegas-Ch, W. (2025). Machine learning models for academic performance prediction: Interpretability and application in educational decision-making. *Frontiers in Education*, 10, 1632315. <https://doi.org/10.3389/feduc.2025.1632315>
10. Heikal, A. F., & Khalil, Z. A. (2023). Student performance prediction using machine learning: A comprehensive analysis. *ASRIC Journal on Engineering Sciences*, 4(1), 266–276.

11. Ismail, L., Materwala, H., & Hennebelle, A. (2021). Comparative analysis of machine learning models for students' performance prediction. In *Advances in digital science* (pp. 149–160). Springer. https://doi.org/10.1007/978-3-030-71782-7_14
12. Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and explainable AI. *Education and Information Technologies*, 27(9), 1–35. <https://doi.org/10.1007/s10639-022-11120-6>
13. Lünich, M., & Keller, B. (2024). 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 (FAccT '24)* (pp. 1031–1042). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3658953>
14. nr, R., Retna, M., & Biji, C. (2024). Explainable machine learning prediction for the academic performance of deaf scholars. *IEEE Access*, 12, 23595–23612. <https://doi.org/10.1109/ACCESS.2024.3363634>
15. Oppong, S. (2023). Predicting students' performance using machine learning algorithms: A review. *Asian Journal of Research in Computer Science*, 16(3), 128–148. <https://doi.org/10.9734/AJRCOS/2023/v16i3351>
16. Soyoye, T., Chen, T., Hill, R., & McCabe, K. (2023). Predicting academic performance of university students using machine learning: A case study in the UK. In *2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (pp. 431–434). IEEE. <https://doi.org/10.1109/WI-IAT59888.2023.00071>