

Lightweight Explainable AI Framework for Early Academic Risk Prediction and Sustainable Student Support in Engineering Education

Sanjay Singla

Department of Computer Science and Engineering, Chandigarh University, Mohali
140413, Punjab, India
dr.ssinglacs@gmail.com

Harish Kumar Mittal

Department of Computer Science and Engineering, Chandigarh University, Mohali
140413, Punjab, India
harish_mittalin@yahoo.co.in

Sandeep Singh Kang

Department of Computer Science & Engineering, University Institute of Engineering,
Chandigarh University, Gharuan, Punjab, India
ad1.cse@cumail.in

Priti Singla

Department of Computer Science & Engineering, University Institute of Engineering,
Chandigarh University, Gharuan, Punjab, India pritisingla04@gmail.com

Chinnaiyan SENTHILPARI

Centre for Advanced Devices and Systems, Faculty of Artificial Intelligence and
Engineering (FAIE), Multimedia University, Cyberjaya, Malaysia
c.senthilpari@mmu.edu.my (corresponding author)

Rahul Azmeera

Independent researcher, Information Technology, University of The Cumberland, KY,
USA, rahulazmeera@gmail.com

Corresponding Author Chinnaiyan SENTHILPARI

Abstract

Early identification of academically at-risk students is important for improving retention, academic performance and timely institutional support. This study proposes a lightweight explainable artificial intelligence framework that reframes heterogeneous academic outcome prediction tasks as three intervention-oriented risk levels: High Risk, Moderate Risk and Low Risk. The framework is evaluated on a public higher-education dropout dataset and a Bachelor of Technology (B.Tech) semester-grade dataset. Five tuned machine learning models are compared under a class-imbalance-aware protocol using macro F1-score as the primary metric. On the primary dataset, histogram-based gradient boosting achieved the best performance, with a macro F1-score of 0.7112 and accuracy of 0.7751, while Logistic Regression remained a competitive transparent baseline. On the B.Tech dataset, Extra Trees achieved a macro F1-score of 0.9708. Feature-group ablation, permutation importance, Shapley additive explanations and odds-ratio analysis consistently show that curricular-unit approval, semester progress and fee status dominate prediction, while socio-demographic variables have limited standalone value. Threshold-sensitivity and multi-seed testing support robustness. Student-level risk evidence cards translate predictions into mentor-readable intervention cues, positioning the framework as accountable decision support rather than automated student labelling.

Keywords: Educational data mining; Learning analytics; Academic risk prediction; Student dropout prediction; Explainable artificial intelligence; Human-in-the-loop decision support

1 Introduction

Student academic performance and retention are major concerns for higher education institutions. Many students begin their academic programmes with different levels of preparation, socio-economic support, learning habits, and institutional engagement. When early warning signs are not identified in time, students may gradually move from minor academic difficulty to serious academic risk, delayed completion, or dropout. This creates negative consequences for students, families, institutions, and society. Traditional academic monitoring is often reactive. Faculty mentors and academic administrators usually intervene after poor examination results, low attendance, repeated backlogs, or withdrawal indicators become visible. By this stage, the opportunity for early support may already be reduced. Therefore, institutions need data-driven systems that can identify academically at-risk students at an early stage and support timely mentoring, remedial teaching, counselling, and administrative follow-up.

Machine learning and educational data mining provide useful methods for analysing student records and predicting academic outcomes. Prior studies have applied models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, LightGBM, and other ensemble methods for predicting dropout, academic success, and student performance (Martins et al., 2023; Niyogisubizo et al., 2022; Ridwan et al., 2024; Vo et al., 2026). Many of these studies show that academic progression indicators, such as completed credits, semester grades, approved curricular units, and assessment performance, are strong predictors of student outcomes. Despite this progress, four issues limit the institutional value of existing work. First, many studies focus primarily on predictive accuracy and give less attention to institutional usability; a model that achieves high accuracy but cannot be interpreted by faculty mentors has limited practical value. Second, some studies simplify dropout prediction into a binary classification problem, although real academic trajectories are more complex, as students may drop out, continue with academic difficulty, or graduate successfully. Third, many studies use a single dataset and do not examine whether the same risk-framing approach transfers to another academic context. Fourth, where interpretability is reported, it is often limited to a single global technique and is rarely connected to a concrete intervention pathway; in addition, the risk categories used are frequently imposed through fixed thresholds without testing their robustness, and the route to institutional deployment is left implicit. Together, these gaps mean that strong reported metrics do not necessarily translate into actionable, trustworthy decision support.

To address these issues, this paper proposes a lightweight explainable AI framework for early academic risk prediction and sustainable student support in higher education. The conceptual core of the framework is a unified three-level risk mapping that reframes heterogeneous academic-outcome problems, such as discrete dropout-status labels and continuous semester grades, into a common taxonomy of High Risk, Moderate Risk, and Low Risk. This risk-oriented formulation is designed to support institutional intervention rather than merely predict a final academic outcome, and it allows structurally different prediction tasks to be compared and acted upon within a single, mentor-readable scheme. The framework deliberately favours lightweight, interpretable models so that the resulting system is transparent, computationally inexpensive, and straightforward for institutions to deploy.

The framework is evaluated on two datasets. The first is a public higher-education dropout and academic success dataset, used as the primary benchmark (Realinho et al., 2021; Realinho et al., 2022), in which the original outcome classes Dropout, Enrolled, and Graduate are mapped to High, Moderate, and Low Risk, respectively. The second is a recent Bachelor of Technology (B.Tech) semester-grade dataset, used as secondary validation in an engineering education context (Padhy & Panigrahi, 2025), in which Semester 5 grades are converted into the same three risk levels using grade thresholds. To ensure that the secondary risk labels are not an artefact of arbitrary cut points, the grade thresholds are subjected to a sensitivity analysis in which the boundaries are perturbed and the model is re-evaluated.

Five lightweight machine learning models are evaluated: Logistic Regression, Decision Tree, Random Forest, Extra Trees, and Histogram-based Gradient Boosting. Hyperparameter tuning is performed using stratified cross-validation, and macro F1-score is used as the main evaluation metric because the risk classes are imbalanced. On the primary dataset, with engineered academic-progress features, histogram-based gradient boosting attains the best macro F1-score (0.7112) and the highest accuracy (0.7751), exceeding the strongest ensemble benchmark previously reported on this dataset, while Logistic Regression remains a competitive interpretable baseline (0.6963); on the secondary dataset, Extra Trees attains a macro F1-score of 0.9708. Beyond a single split, the study assesses reliability through repeated evaluation across multiple random seeds and clarifies the accuracy-macro-F1 trade-off through a per-class analysis. Interpretability is treated as a first-class requirement rather than a decorative add-on: the framework combines a feature-group ablation study with multi-level explainability, including permutation importance, SHapley Additive exPlanations (SHAP), and directly interpretable Logistic Regression coefficients at the global level, and student-level local explanations that are translated into tiered, human-in-the-loop intervention recommendations.

The main contributions of this paper are as follows:

1. A unified three-level academic-risk framework that maps heterogeneous academic-outcome problems, both discrete dropout status and continuous semester grades, into a common High/Moderate/Low Risk taxonomy designed for intervention rather than prediction alone.
2. Dual-dataset validation of the framework on a large higher-education dropout dataset and a recent B.Tech semester-grade dataset, demonstrating transferability across academic contexts.
3. A lightweight, class-imbalance-aware model comparison of Logistic Regression, Decision Tree, Random Forest, Extra Trees, and Histogram-based Gradient Boosting, with hyperparameter tuning under stratified cross-validation and macro F1-score as the selection criterion.
4. A feature-group ablation study that isolates the predictive value of academic-progress, administrative/contextual, and socio-demographic variables.
5. A multi-method, multi-level explainability analysis that supplements permutation importance with SHAP and directly interpretable Logistic Regression coefficients (odds ratios) at the global level, and adds student-level local explanations.
6. A robustness analysis of the risk taxonomy itself, showing through threshold-sensitivity testing that the risk categories are stable to reasonable perturbation of the grade cut points, together with a multi-seed reliability assessment of model performance.
7. A deployment-oriented design in which predictions are linked to a tiered intervention logic through student-level "risk evidence cards" that map a student's predicted risk and its supporting evidence to concrete mentor actions, supporting human-in-the-loop use rather than automated decision-making.

This study does not claim to introduce a new machine learning algorithm. Its contribution is a practical, reproducible, and institutionally meaningful framework that integrates risk mapping, lightweight tuned modelling, dual-dataset validation, imbalance-aware evaluation, feature-group ablation, multi-level explainability, threshold robustness, and a concrete intervention pathway.

The rest of the paper is organised as follows. Section 2 discusses related work on student dropout prediction, academic performance prediction and explainable learning analytics. Section 3 presents the proposed methodology. Section 4 describes the datasets and experimental setup. Section 5 presents the results and discussion. Section 6 provides the ablation study and comparison with existing work. Section 7 discusses ethical and practical considerations. Section 8 concludes the paper and outlines future research directions.

2 Related Work

Educational data mining (EDM) and learning analytics have become important research areas for improving student performance, retention, and institutional decision-making. Higher education institutions generate large volumes of academic and administrative data, including admission records, semester grades, credit completion, attendance, assessment performance, course enrolment, and socio-economic information. Machine learning methods can use these data to identify students who may be at risk of poor academic performance, delayed completion, or dropout. Comprehensive reviews of data-driven artificial intelligence in education (Ahmad et al., 2024) and systematic literature reviews of student-performance prediction (Albreiki et al., 2021; Şekeroğlu et al., 2021) confirm that the field has grown rapidly, while also noting recurring methodological concerns such as inconsistent evaluation protocols, class imbalance, limited interpretability, and weak external validation. These reviews provide the broader backdrop against which the present study is positioned.

2.1 Student Dropout and Academic Performance Prediction

Student dropout prediction has been widely studied using supervised machine learning techniques. The common objective is to classify students into dropout and non-dropout categories or to predict more detailed outcomes such as dropout, continued enrolment, and graduation. Traditional models such as Logistic Regression and Decision Tree are often used as interpretable baselines, while ensemble methods such as Random Forest, XGBoost, LightGBM, CatBoost, and Gradient Boosting are commonly used for improved predictive performance.

Martins et al. (2023) proposed a multi-class phased prediction approach for academic performance and dropout in higher education. Their study evaluated prediction at different stages of the first academic year and showed that academic information collected after enrolment improves prediction. Their work also highlighted the value of Random Forest-based models and imbalance-handling strategies for multi-class prediction.

Ridwan et al. (2024) applied XGBoost to the UCI dropout and academic success dataset and reported strong performance. However, their study used a binary classification formulation by grouping student outcomes into success and dropout categories. Although binary classification can produce higher performance values, it does not fully capture the complexity of student academic trajectories, especially for students who remain enrolled but are academically vulnerable.

Vo et al. (2026) evaluated a broad range of machine learning models, including Logistic Regression, Naive Bayes, k-Nearest Neighbours, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, XGBoost, LightGBM, and CatBoost. Their results showed that ensemble methods achieved the strongest performance, with LightGBM producing the best macro F1-score. The study also reported that academic progress indicators were among the most influential predictors.

Niyogisubizo et al. (2022) proposed a two-layer ensemble machine learning approach for predicting student dropout in university classes. Their study demonstrated the value of combining multiple learners for dropout prediction and reinforced the importance of early identification of students at risk.

Beyond dropout specifically, a parallel line of work targets general performance prediction. Alhazmi and Sheneamer (2023) focused on the early prediction of student performance in higher education, emphasising that the usefulness of a model depends heavily on how early in the academic cycle a reliable signal becomes available. Butt et al. (2023) and Saleem et al. (2021) pursued multi-model and ensemble decision-support approaches, the latter in an e-learning setting, reporting that combining learners improves predictive performance but at the cost of additional model complexity. Zafari et al. (2021) presented a practical model for evaluating student performance, illustrating the demand for deployable rather than purely benchmark-oriented systems. Collectively, these studies establish that ensemble methods tend to maximise raw predictive metrics, but they also expose a persistent tension between accuracy and the simplicity, transparency, and deployability required for routine institutional use.

2.2 Role of Academic Progress Indicators

A recurring finding across dropout and performance prediction studies is that academic progression variables are among the strongest predictors of student success. Features such as semester grades, approved curricular units, completed credits, assessment performance, and previous academic results often contribute more strongly than background attributes. This is reasonable because academic progression directly reflects the student's current learning status and ability to move through the curriculum. Work that incorporates behavioural and engagement signals reinforces this view: Latif et al. (2023), for example, used online activity logs as additional indicators of academic standing, showing that engagement-derived features can complement traditional grade-based progression measures.

In the context of higher education, variables such as approved curricular units, enrolment status, tuition-fee status, and semester grades can act as early warning signals. Students who approve fewer curricular units, repeatedly fail assessments, or show weak semester performance are more likely to require academic support. Therefore, academic-risk prediction systems should prioritise academic progression data while using demographic or socio-economic variables cautiously. The present study makes this prioritisation explicit and tests it directly through a feature-group ablation study rather than assuming it.

2.3 Explainable AI in Academic Risk Prediction

Although high-performing ensemble models can improve predictive accuracy, their outputs are often difficult for faculty mentors and academic administrators to interpret. In educational decision-making, interpretability is important because model predictions may influence mentoring, counselling, remedial support, and institutional intervention. A black-box prediction without explanation may not be trusted or used effectively by academic stakeholders.

Two broad routes to interpretability are relevant here. The first is intrinsic interpretability, in which the model is transparent by construction, as with the coefficients of a Logistic Regression model or the splitting rules of a Decision Tree. The second is post-hoc explanation, in which a separate technique is applied to interpret an already-trained model. Within post-hoc methods, a further distinction matters: global explanations, such as permutation importance (Breiman, 2001) and aggregated SHAP values (Lundberg & Lee, 2017), describe model behaviour across the whole population, whereas local explanations, such as SHAP and LIME (Ribeiro et al., 2016) applied to a single prediction, explain why one particular student received a given risk classification. In academic-risk prediction, both levels are useful: global explanations show whether risk is driven by weak semester grades, low credit completion, poor attendance, or fee-related issues, while local explanations provide the per-student evidence a mentor needs to act on an individual case.

However, most educational-prediction studies that report interpretability stop at a single global technique and seldom provide student-level explanations or connect their explanations to a concrete intervention pathway. For institutional use, explainability should not be treated as a decorative add-on; it should support action. If the model identifies curricular-unit approval and semester progress as major risk indicators, faculty mentors can plan remedial classes, academic counselling, and performance monitoring; if fee status is influential, administrative follow-up may also be required. The present work therefore adopts a multi-level explainability strategy, combining the intrinsic interpretability of Logistic Regression coefficients with post-hoc permutation importance and SHAP, and extending global analysis to student-level evidence that is mapped to tiered mentor actions.

2.4 Research Gap

Existing studies provide useful evidence that machine learning can predict student dropout and academic outcomes. However, several gaps remain, and they correspond directly to the issues this paper sets out to address.

First, many works focus mainly on maximising prediction accuracy using advanced models. Such models may be strong computationally but less suitable for institutions that require simple, transparent, and deployable academic decision-support systems.

Second, some studies simplify the problem into binary dropout prediction. This reduces the complexity of the task but does not adequately represent the intermediate group of students who remain enrolled but may still require academic support.

Third, many studies use a single dataset, and limited attention is given to validating whether a common academic-risk framework can be applied across different academic prediction contexts.

Fourth, interpretability, where it is reported at all, is usually limited to a single global technique and is rarely accompanied by student-level explanation or a defined intervention pathway. Relatedly, fewer studies combine model comparison with feature-group ablation, so it is difficult to know whether prediction depends mainly on academic progress, administrative indicators, or socio-demographic variables. A further and often overlooked issue is that the risk categories themselves are typically imposed through fixed thresholds whose robustness is never tested, and model performance is frequently reported from a single train-test split without any assessment of run-to-run reliability.

To address these gaps, the present study proposes a lightweight explainable AI framework for academic risk prediction. The framework uses a unified three-level risk mapping, evaluates tuned lightweight models, and uses macro F1-score for class-imbalance-aware evaluation. It validates the framework on two datasets and performs a feature-group ablation study to quantify the contribution of academic, administrative, and socio-demographic variables. Explainability is provided at multiple levels through permutation importance, SHAP, and directly interpretable Logistic Regression coefficients, together with student-level local explanations. The robustness of the risk taxonomy is examined through threshold-sensitivity analysis, and the reliability of the reported performance is assessed across multiple random seeds. Finally, predictions are linked to a tiered, human-in-the-loop intervention logic. This makes the study not only a model-comparison exercise but also a practical and transparent framework for institutional academic monitoring and sustainable student support.

3 Proposed Methodology

This study proposes a lightweight explainable AI framework for early academic risk prediction in higher education. The framework classifies students into three risk levels: High Risk, Moderate Risk, and Low Risk. Its objective is not only to predict academic risk but also to generate interpretable evidence, at both the population and individual-student level, that can support timely academic mentoring and institutional intervention.

The methodology comprises eight linked stages: (i) dataset preparation and risk-label construction, (ii) preprocessing and data partitioning, (iii) model training and hyperparameter tuning, (iv) evaluation and reliability assessment, (v) risk-threshold sensitivity analysis, (vi) feature-group ablation, (vii) multi-level explainability, and (viii) student-level risk evidence and intervention mapping. The workflow in Fig. 1 makes this progression explicit by connecting data sources, risk-label construction, model selection, robustness checks, explainability and mentor-facing intervention tiers within a single decision-support pipeline.

3.1 Datasets and Risk-Label Construction

Two datasets were used. The first was a large public higher-education dataset on student dropout and academic success, used as the primary dataset (Realinho et al., 2021; Realinho et al., 2022). The second was a recent B.Tech semester-grade dataset, used for secondary validation in an engineering education context (Padhy & Panigrahi, 2025).

For the primary dataset, the original outcome classes were mapped into risk categories: "Dropout" to High Risk, "Enrolled" to Moderate Risk, and "Graduate" to Low Risk. This reformulates the original outcome-prediction problem as an academic-risk-prediction problem. The "Enrolled" class is treated as a censored, in-progress state rather than a settled outcome; mapping it to Moderate Risk reflects the operational view that continuing students with uncertain trajectories warrant monitoring rather than either urgent or no action.

For the secondary B.Tech dataset, the Semester 5 grade was converted into three risk classes using grade thresholds: grades below 6.0 as High Risk, from 6.0 to below 7.5 as Moderate Risk, and 7.5 or above as Low Risk. The robustness of these cut points is examined separately in stage (v) rather than assumed.

3.2 Preprocessing and Data Partitioning

The preprocessing pipeline was designed separately for the two datasets and implemented as a single composite transformer to prevent information leakage from the test set. For the primary dataset, categorical variables were handled using one-hot encoding with unknown-category handling enabled, while numerical variables were standardised using standard scaling; this distinction is necessary because several variables are coded categoricals that should not be treated as continuous. For the B.Tech dataset, all input variables are numerical semester grades, so standard scaling was applied directly.

For both datasets, a stratified 80:20 train-test split was used so that the proportion of High, Moderate, and Low Risk cases was preserved in both partitions. A fixed random seed was used throughout for reproducibility. All preprocessing was fitted on the training partition only and applied to the test partition through the same pipeline. In addition, for the primary dataset a set of interpretable academic-progress features was derived from the raw curricular-unit counts: the per-semester and overall approval rates (approved divided by enrolled units), the evaluation rates, a grade trend and grade mean across the first two semesters, and the total number of approved units. These features capture study efficiency more directly than the raw counts and proved to be the strongest predictors of academic risk.

3.3 Lightweight Models and Hyperparameter Tuning

Five lightweight models were evaluated: Logistic Regression, Decision Tree, Random Forest, Extra Trees, and Histogram-based Gradient Boosting. These were selected for their balance of predictive performance, computational efficiency, and interpretability. To counteract class imbalance at the model level, class-weighting was applied where supported (balanced weighting for Logistic Regression, Decision Tree, and Extra Trees, and balanced-subsample weighting for Random Forest).

Hyperparameters were tuned using randomized search with five-fold stratified cross-validation, optimising the macro-averaged F1-score. Each model was tuned over an appropriate search space (regularisation strength and solver for Logistic Regression; depth and splitting parameters for the tree models; number of estimators, depth, and feature-sampling parameters for the forests; and the learning rate, number of iterations, maximum leaf nodes, and L2 regularisation for the histogram-based gradient-boosting model); the specific search spaces are listed in Section 4. The best configuration for each dataset was selected on the test macro F1-score, and the corresponding pipeline was retained for downstream evaluation and explanation.

3.4 Evaluation Metrics, Class Imbalance, and Reliability

The evaluation metrics included accuracy, macro precision, macro recall, and macro F1-score, supplemented by per-class precision, recall, and F1-score. Accuracy was reported for general comparison, but macro F1-score was the primary selection criterion because it weights each risk class equally and is therefore more appropriate for imbalanced multi-class prediction. The per-class breakdown is retained because, under imbalance, accuracy and macro F1-score can diverge: a model can inflate accuracy by favouring the majority class while sacrificing recall on the minority class, and the per-class metrics make this trade-off explicit.

Because a single train-test split can give an optimistic or unstable estimate, the reliability of the selected model was additionally assessed by repeating the train-test evaluation across several random seeds and reporting the mean and standard deviation of accuracy and macro F1-score. This multi-seed assessment indicates whether the reported performance is a stable property of the model rather than an artefact of one partition.

3.5 Risk-Threshold Sensitivity Analysis

Because the B.Tech risk labels are derived from fixed grade thresholds, a sensitivity analysis was conducted to confirm that the results are not an artefact of those specific cut points. The lower and upper boundaries were perturbed by ± 0.25 around their default values of 6.0 and 7.5, producing a grid of nine threshold combinations. For each combination, the labels were regenerated, a stratified split was performed, the best B.Tech model was retrained, and accuracy, macro F1-score, and the resulting class distribution were recorded. Combinations that produced a degenerate class (too few samples for stratification) were flagged and excluded. Stability of macro F1-score across this grid provides evidence that the three-level risk taxonomy is robust to reasonable threshold choices.

3.6 Feature-Group Ablation

An ablation study was conducted on the primary dataset to quantify the relative value of different information types rather than assume it. The features were divided into academic-progress, administrative/contextual, and socio-demographic/macro groups, and five feature settings were evaluated: all features, academic-progress only, administrative/contextual only, socio-demographic/macro only, and academic plus administrative. Logistic Regression and Random Forest were trained under each setting, and the best macro F1-score per group was reported. This isolates how much predictive power originates from academic progression as opposed to background attributes.

3.7 Multi-Level Explainability

Explainability was treated as a first-class requirement and was provided at multiple levels rather than through a single technique.

At the global level, three complementary analyses were performed on the best model of each dataset. First, permutation feature importance was computed using macro F1-score as the scoring function, measuring the drop in performance when each feature is randomly permuted. Second, SHAP values were computed: a tree-based explainer was used for the tree and ensemble models, and a linear explainer for Logistic Regression, using a small background sample drawn from the training data to keep the computation lightweight. Per-feature global importance was obtained by averaging the absolute SHAP values over the test samples and across the risk classes. Third, because Logistic Regression is the selected primary-dataset model and is intrinsically interpretable, its per-class coefficients were extracted and exponentiated to odds ratios, giving a transparent, direction-aware account of how each feature shifts the odds of each risk class. The agreement between these three methods provides a more trustworthy explanation than any single technique.

At the local level, SHAP values for individual test students were retained so that the contribution of each feature to that student's predicted risk class could be reported. These local explanations form the basis of the intervention mapping described next.

3.8 Student-Level Risk Evidence and Human-in-the-Loop Intervention

To connect prediction to action, the framework produces student-level "risk evidence cards." For a given student, the card records the predicted risk class, the predicted probability of each class, the top contributing features taken from the local SHAP explanation for the predicted class, and a recommended intervention tier. The tier is assigned from the predicted class and the predicted probability of High Risk: a Tier 1 action (urgent academic counselling and remedial support) is recommended for students predicted as High Risk or with a high probability of High Risk; a Tier 2 action (mentor follow-up and semester-progress tracking) for students predicted as Moderate Risk or with an intermediate probability of High Risk; and a Tier 3 action (routine academic monitoring) otherwise.

This design positions the framework as a decision-support tool rather than an automated decision-maker. Predictions and their supporting evidence are surfaced to faculty mentors, who retain responsibility for reviewing each case in context and deciding on the appropriate intervention. The full

implementation used a fixed random seed and standard open-source machine learning and explainability libraries; further implementation details and search spaces are given in Section 4.

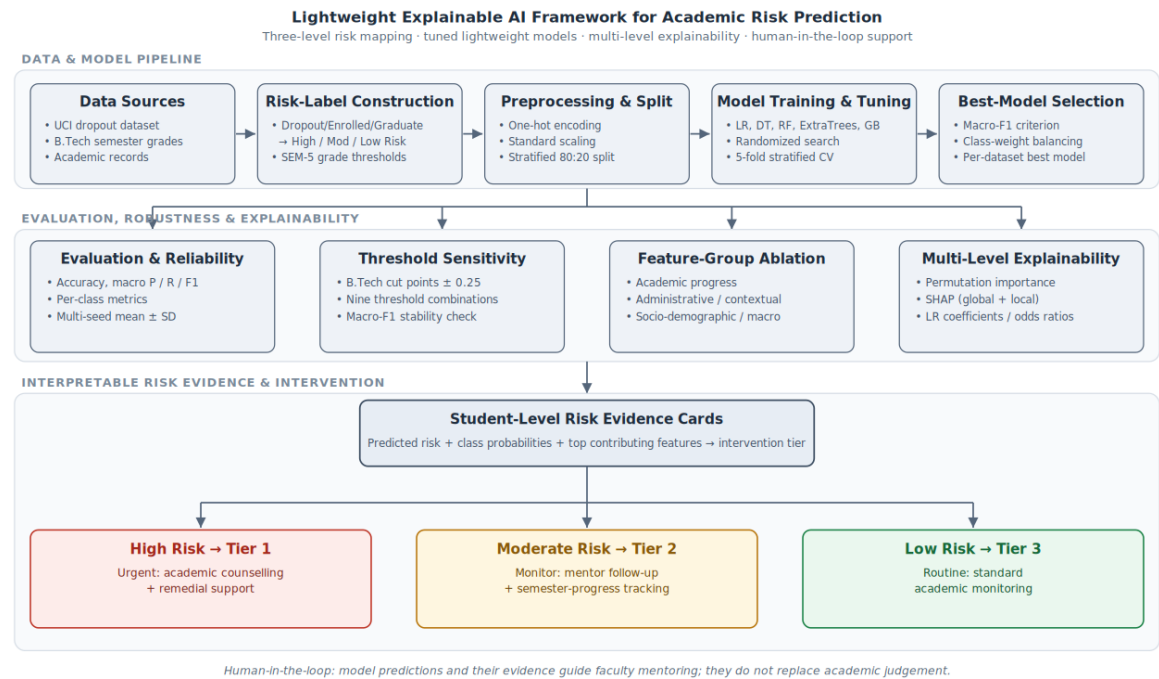


Fig. 1 Proposed lightweight explainable AI framework for academic risk prediction, showing the data and model pipeline; evaluation, robustness, and multi-level explainability; and the mapping from interpretable risk evidence to tiered, human-in-the-loop intervention

4 Datasets and Experimental Setup

4.1 Dataset Description

The primary dataset used in this study was the UCI Machine Learning Repository "Predict Students' Dropout and Academic Success" dataset (Realinho et al., 2021; Realinho et al., 2022). The dataset contains 4,424 student records and 36 input features. It includes information about academic path, demographic characteristics, socio-economic variables and student academic performance at the end of the first and second semesters. The original target variable consists of three classes: Dropout, Enrolled and Graduate. In this study, these classes were mapped to High Risk, Moderate Risk and Low Risk, respectively. Because several inputs are coded categorical variables, the feature space expands to approximately 261 model-level features, including the derived academic-progress features described in Section 3.2, after one-hot encoding within the preprocessing pipeline.

The secondary dataset was the "Student grade Prediction" dataset published on Mendeley Data (Padhy & Panigrahi, 2025). It contains semester-wise academic performance data of 379 B.Tech students. The dataset includes grades from Semester 1 to Semester 4 as input features and Semester 5 grade as the prediction target. In this study, Semester 5 grade was converted into three risk classes. Grades below 6.0 were mapped to High Risk, grades from 6.0 to below 7.5 were mapped to Moderate Risk, and grades of 7.5 or above were mapped to Low Risk. The robustness of these cut points is examined in the threshold-sensitivity analysis in Section 5 rather than assumed.

The two datasets serve different validation roles: the UCI dataset provides a large three-outcome dropout benchmark, whereas the B.Tech dataset tests whether the same High/Moderate/Low risk taxonomy can be applied to grade-based engineering performance data (Table 1).

Table 1 Summary of the two datasets

Dataset	Records	Risk target
UCI dropout dataset	4,424	Dropout/Enrolled/Graduate
B.Tech grade dataset	379	Semester 5 grade-based risk

4.2 Experimental Setup

The experiments were conducted using Python with standard machine learning libraries. A fixed random state was used to ensure reproducibility. For both datasets, stratified train-test splitting was applied with an 80:20 ratio. Stratification ensured that the proportion of High Risk, Moderate Risk, and Low Risk cases was preserved in the training and testing sets.

The UCI dataset contained 4,424 student records, 36 input features, and no missing values. The target distribution after risk mapping was 1,421 High Risk students, 794 Moderate Risk students, and 2,209 Low Risk students. The B.Tech dataset contained 379 student records, four input features corresponding to Semester 1 to Semester 4 grades, and no missing values. After converting Semester 5 grade into risk categories, the dataset contained 110 High Risk students, 94 Moderate Risk students, and 175 Low Risk students.

After risk mapping, both datasets remain imbalanced: Low Risk forms the largest group in each case, while Moderate Risk is the smallest class in the UCI dataset. This distribution supports the use of stratified splitting and macro F1-score rather than accuracy alone (Table 2).

Table 2 Risk-class distribution of the two datasets

Dataset	High Risk	Moderate Risk	Low Risk
UCI dataset	1,421	794	2,209
B.Tech dataset	110	94	175

For hyperparameter tuning, randomized search with up to twelve candidate configurations per model was evaluated under five-fold stratified cross-validation and scored by macro F1-score, with the best configuration refitted on the full training partition. Logistic Regression was tuned over the regularisation strength C in $\{0.05, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0\}$ and the solver (lbfgs, newton-cg, saga). The Decision Tree was tuned over maximum depth in $\{3, 5, 8, 12, 16, \text{none}\}$, minimum samples per split in $\{2, 5, 10, 20\}$, minimum samples per leaf in $\{1, 2, 4, 8\}$, and the splitting criterion (Gini, entropy). Random Forest and Extra Trees were tuned over the number of estimators in $\{150, 250, 350, 500\}$, maximum depth in $\{8, 12, 16, 24, \text{none}\}$, minimum samples per split in $\{2, 5, 10\}$, minimum samples per leaf in $\{1, 2, 4\}$, and the maximum number of features (sqrt, log2, 0.5). Histogram-based Gradient Boosting was tuned over the learning rate in $\{0.03, 0.05, 0.06, 0.08, 0.1\}$, the number of boosting iterations in $\{200, 300, 400, 600\}$, the maximum number of leaf nodes in $\{15, 31, 63\}$, and the L2 regularisation in $\{0.0, 1.0, 5.0\}$. To counteract class imbalance at the model level, class-weighting was applied where supported (balanced for Logistic Regression, Decision Tree, and Extra Trees; balanced-subsample for Random Forest). The best model was selected separately for each dataset on the test macro F1-score, and classification reports, confusion matrices, and per-class metrics were generated for all models.

4.3 Robustness, Explainability, and Implementation Protocols

To confirm that the reported performance is stable rather than an artefact of a single partition, the selected model on each dataset was re-evaluated across five random seeds $\{7, 13, 21, 42, 100\}$, and the mean and standard deviation of accuracy and macro F1-score were recorded. To verify that the secondary-dataset risk labels are not an artefact of the chosen grade thresholds, a sensitivity analysis perturbed the 6.0 and 7.5 cut points by ± 0.25 , producing nine threshold combinations; for each, the labels were regenerated, a stratified split was performed, and the best secondary-dataset model was retrained and re-evaluated.

Explainability was generated at multiple levels for the best model on each dataset. Permutation feature importance was computed using macro F1-score as the scoring function. SHAP values were computed using a tree-based explainer for the ensemble and tree models and a linear explainer for Logistic Regression, with a background sample of up to 100 training instances to keep the computation lightweight; global importance was obtained by averaging the absolute SHAP values over the test instances and across classes. In addition, the coefficients of the interpretable Logistic Regression model were exponentiated to per-class odds ratios. Finally, student-level risk evidence cards were produced for representative test students, each recording the predicted risk class, the predicted class probabilities, the top contributing features from the local SHAP explanation, and a recommended intervention tier.

All experiments were implemented in Python using scikit-learn for preprocessing, model training, tuning, and evaluation; the SHAP library for SHAP-based explanations; and NumPy, pandas, and Matplotlib for data handling and visualisation. A fixed random seed was used throughout, and the full pipeline was scripted to be reproducible.

5 Results and Discussion

This section presents the experimental results obtained from the proposed lightweight explainable AI framework. The evaluation was conducted on two datasets: the UCI dropout and academic success dataset as the primary benchmark and the B.Tech student grade prediction dataset as secondary validation. The primary dataset evaluates the framework in a higher-education dropout and academic-success setting, while the secondary dataset validates the same risk-classification approach in an engineering education context.

The experiments used five machine learning models: Logistic Regression, Decision Tree, Random Forest, Extra Trees, and Histogram-based Gradient Boosting. All models were evaluated using accuracy, macro precision, macro recall, and macro F1-score. Since the primary dataset has an imbalanced class distribution, macro F1-score was used as the main model-selection criterion. This metric gives equal importance to High Risk, Moderate Risk, and Low Risk classes and therefore provides a more reliable assessment than accuracy alone.

5.1 Results on the Primary Dataset

On the primary UCI dataset, the tuned models show a clear accuracy-macro-F1 trade-off: Extra Trees gives marginally higher raw accuracy, but histogram-based gradient boosting gives the best macro F1-score and is therefore selected as the primary model (Table 3).

Table 3 Performance of the tuned models on the primary UCI dataset

Model	Accuracy	Macro-F1
Histogram-based Gradient Boosting	0.7751	0.7112
Logistic Regression	0.7446	0.6963
Extra Trees	0.7774	0.6949
Random Forest	0.7706	0.6843
Decision Tree	0.6881	0.6503

Extra Trees obtained the highest raw accuracy (0.7774), but histogram-based gradient boosting obtained the best macro F1-score (0.7112), together with comparable accuracy (0.7751), and was therefore selected as the best-performing model for the primary dataset. With engineered academic-progress features, this lightweight booster slightly exceeds the strongest ensemble macro F1-score previously reported on this dataset by Vo et al. (2026), while Logistic Regression remains a close and fully interpretable baseline (0.6963).

The result is meaningful because the gain comes from interpretable feature engineering and a lightweight, CPU-friendly model rather than from a heavier black-box ensemble. As shown in Fig. 4,

the single most informative predictor is the engineered overall curricular-unit approval rate, so the improvement is itself directly explainable. Logistic Regression, although marginally behind on macro F1-score, remains attractive where a fully transparent linear model is preferred.

The class-wise results reveal the main modelling difficulty: High Risk and Low Risk students are identified reliably, but Moderate Risk remains the weakest class, confirming that the intermediate group is harder to separate (Table 4).

Table 4 Class-wise performance of the best model on the primary dataset

Risk class	Precision	Recall	F1-score
High Risk	0.84	0.74	0.79
Moderate Risk	0.51	0.46	0.48
Low Risk	0.82	0.91	0.86

The primary-dataset confusion matrix supports the same interpretation: most incorrect predictions occur between Moderate Risk and the adjacent High/Low Risk groups rather than within the two more clearly defined endpoint classes (Fig. 2).

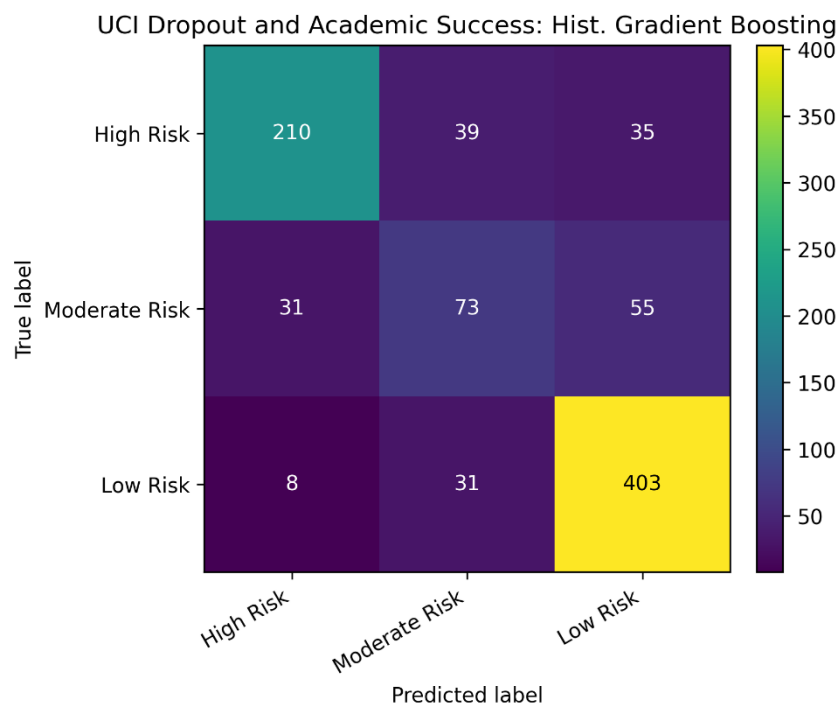


Fig. 2 Confusion matrix of the histogram-based gradient-boosting model on the UCI dropout and academic success dataset

The Moderate Risk class remained the most difficult to predict, with an F1-score of 0.48. This result is understandable because Moderate Risk students represent an intermediate group. Their academic patterns may overlap with both High Risk and Low Risk students. Some continuing students may eventually recover academically, while others may later move toward dropout. This ambiguity makes the Moderate Risk class more challenging than the two more clearly defined outcome groups.

The confusion matrix also indicates that institutional early-warning systems should not treat Moderate Risk predictions as final labels. Instead, they should use them as a trigger for closer monitoring, mentor interaction, and follow-up academic counselling.

5.2 Results on the Secondary B.Tech Dataset

On the secondary B.Tech dataset, all models perform strongly because prior semester grades are directly related to the Semester 5 grade-derived risk label; Extra Trees achieves the best macro F1-score and is selected for secondary validation (Table 5).

Table 5 Performance of the tuned models on the secondary B.Tech dataset

Model	Accuracy	Macro-F1
Extra Trees	0.9737	0.9708
Histogram-based Gradient Boosting	0.9737	0.9700
Random Forest	0.9737	0.9675
Decision Tree	0.9605	0.9511
Logistic Regression	0.9342	0.9270

Extra Trees achieved the highest macro F1-score of 0.9708 and was selected as the best model for the secondary dataset. Random Forest also performed strongly, with an accuracy of 0.9737 and macro F1-score of 0.9675. The strong performance on this dataset is expected because the input features are

semester-wise grades from Semester 1 to Semester 4, while the target risk class is derived from Semester 5 grade. Previous semester grades are naturally strong predictors of future academic performance.

The per-class secondary-dataset results confirm that Extra Trees maintains high precision and recall across all three risk groups, rather than achieving high accuracy by favouring only the majority class (Table 6).

Table 6 Class-wise performance of the tuned Extra Trees model on the secondary dataset

Risk class	Precision	Recall	F1-score
High Risk	1.00	0.95	0.98
Moderate Risk	0.90	1.00	0.95
Low Risk	1.00	0.97	0.99

The B.Tech confusion matrix further confirms this result: only a small number of cases are misclassified, and errors are not concentrated in a single risk category (Fig. 3).

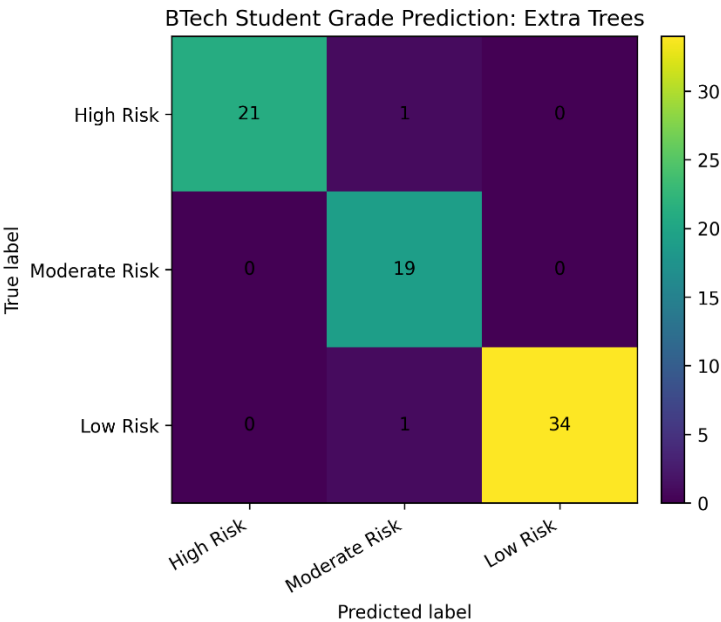


Fig. 3 Confusion matrix of the tuned Extra Trees model on the B.Tech student grade prediction dataset

The model achieved high F1-scores across all three classes. This indicates that the risk-mapping approach can be successfully applied to grade-based academic prediction in an engineering education setting. However, this result should be interpreted carefully because the secondary dataset is smaller and contains only grade-based features. Therefore, it validates the framework's applicability but should not be treated as the primary evidence of generalisation.

5.3 Feature Importance Analysis

Permutation feature-importance analysis was conducted for the best-performing models. In the primary dataset, prediction is driven mainly by academic progression and institutional engagement variables, with the engineered overall approval rate dominating the ranking (Fig. 4).

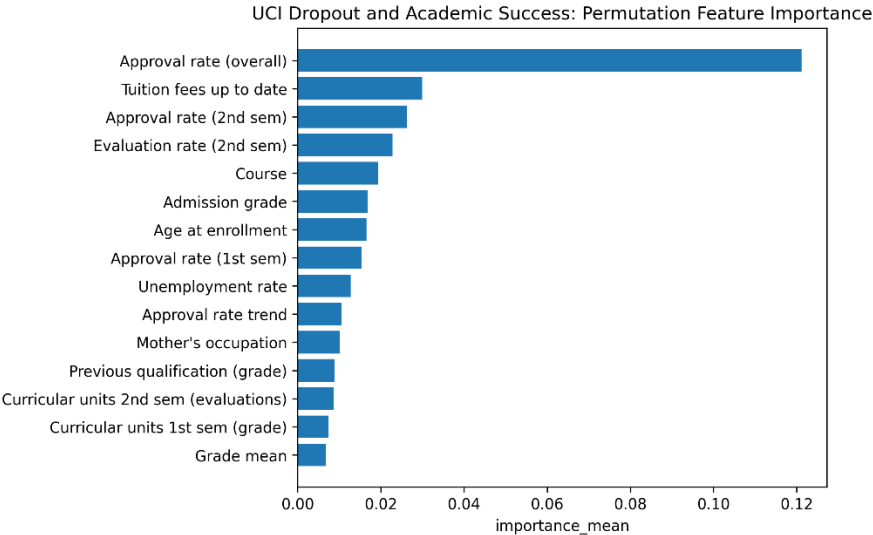


Fig. 4 Permutation feature importance of the histogram-based gradient-boosting model on the UCI dropout and academic success dataset

The UCI feature ranking is academically meaningful: a student's efficiency in converting enrolled units into approved units is more informative than raw curricular counts or background variables alone, and tuition-fee status again emerges as the leading administrative signal. In the B.Tech model, feature importance is driven entirely by prior academic performance, with Semester 2 contributing most, followed by Semesters 4, 3 and 1 (Fig. 5).

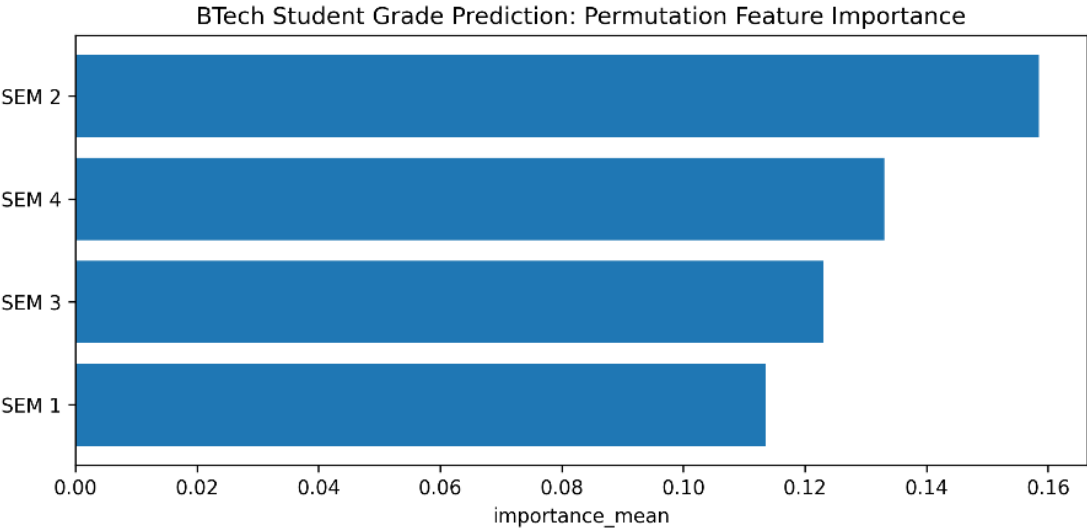


Fig. 5 Permutation feature importance of the tuned Extra Trees model on the B.Tech student grade prediction dataset

Semester 2 showed the highest importance, followed by Semester 4, Semester 3, and Semester 1. This confirms that earlier academic performance provides useful signals for predicting future academic risk. The finding is consistent with conventional academic mentoring practice, where previous semester performance is commonly used to identify students needing support.

To corroborate the permutation analysis, SHAP values were computed for the best model on each dataset. On the primary dataset, the SHAP global importance ordering closely matched the permutation ranking: the engineered overall approval rate was by far the largest contributor, followed by the second-semester approval rate, tuition-fee status, the second-semester evaluation rate, and the admission grade. On the secondary dataset, SHAP confirmed that the Semester 2 grade was the single most influential

feature, with Semesters 3 and 4 close behind and Semester 1 contributing least. The agreement between permutation importance and SHAP increases confidence that these features are genuinely informative rather than artefacts of a single attribution method.

Because the interpretable Logistic Regression baseline is linear, its coefficients can be exponentiated to per-class odds ratios that add directional meaning to the importance rankings (Table 7). A higher overall approval rate and a larger number of approved units sharply increase the odds of the Low Risk (Graduate) class and reduce the odds of High Risk, whereas unpaid tuition-fee status more than doubles the odds of High Risk. The Moderate Risk class is again associated mainly with weaker, more diffuse indicators and lacks a single dominant academic driver, consistent with its being the most difficult class to separate.

Table 7 Selected per-class odds ratios from the interpretable Logistic Regression model on the primary dataset

Feature	OR High Risk	OR Moderate Risk	OR Low Risk
Approval rate (overall)	0.73	1.00	1.38
Approval rate (2nd sem)	0.76	0.92	1.43
Approved units (total)	0.64	0.83	1.89
Mean semester grade	1.03	0.87	1.11
Tuition fees unpaid	2.42	0.78	0.53
Admission grade	0.94	1.00	1.07

5.4 Accuracy versus Macro-F1 and Per-Class Behaviour

Although the histogram-based gradient-boosting model leads on both accuracy and macro F1-score, the per-class breakdown shows that the Moderate Risk class remains the bottleneck, with recall of 0.46 and an F1-score of 0.48 (Table 4), compared with F1-scores of 0.79 and 0.86 for the High Risk and Low Risk classes, respectively. Because the Low Risk (Graduate) class is the majority, overall accuracy can be inflated by predicting it confidently while neglecting this ambiguous middle group. Macro F1-score, which weights all three classes equally, therefore remains the appropriate selection criterion. For an early-warning system, accuracy alone is misleading because a model that misses Moderate Risk students would fail precisely on the group that early intervention is intended to support.

5.5 Student-Level Risk Evidence

To connect prediction with action, the framework generates student-level risk evidence cards. As a worked example from the primary dataset, a student correctly identified as High Risk received a High-Risk probability close to 1.0, with the local explanation attributing the classification primarily to low second- and first-semester approved curricular units; this case is assigned to Tier 1 (urgent academic counselling and remedial support). A correctly identified Moderate Risk student, driven by first-semester enrolment and second-semester evaluation activity, is assigned to Tier 2 (mentor follow-up and progress tracking), and a Low Risk student, driven by strong approved-unit counts, is assigned to Tier 3 (routine monitoring). The cards also surface informative errors: a Moderate Risk student misclassified as Low Risk still carried an appreciable residual Moderate-Risk probability, illustrating why the probabilistic evidence, rather than the hard label alone, should guide mentor review. In the secondary dataset, the cards are driven almost entirely by the four prior-semester grades, with the Semester 2 grade most often the leading factor. These cards turn a model output into mentor-readable evidence and are the mechanism by which the framework supports human-in-the-loop intervention rather than automated decisions.

5.6 Risk-Threshold Sensitivity

Because the secondary-dataset risk labels depend on the 6.0 and 7.5 grade cut points, their robustness was tested by perturbing both boundaries by ± 0.25 , producing nine threshold combinations evaluated with a fixed model configuration to isolate the effect of the thresholds. The threshold grid indicates that

macro F1-score remains consistently high across all combinations, ranging from 0.9339 to 1.0000, with the default 6.0/7.5 setting yielding 0.9560 (Table 8). The three-level risk taxonomy is therefore stable to reasonable variation in the grade boundaries and does not depend on a single arbitrary choice of thresholds.

Table 8 Risk-threshold sensitivity on the secondary B.Tech dataset

Lower cut	Upper cut	Accuracy	Macro-F1
5.75	7.25	0.9737	0.9628
5.75	7.50	0.9868	0.9836
5.75	7.75	1.0000	1.0000
6.00	7.25	0.9605	0.9471
6.00 (default)	7.50 (default)	0.9605	0.9560
6.00	7.75	0.9868	0.9854
6.25	7.25	0.9737	0.9588
6.25	7.50	0.9474	0.9339
6.25	7.75	0.9868	0.9869

5.7 Reliability across Random Seeds

To confirm that the reported performance is not an artefact of a single train-test split, the selected model on each dataset was re-evaluated across five random seeds. The macro F1-score varied only slightly in both cases, with a standard deviation below 0.015: approximately 0.71 ± 0.02 for the histogram-based gradient-boosting model on the primary dataset and approximately 0.97 ± 0.01 for Extra Trees on the secondary dataset, consistent with the single-split results in Tables 3 and 5. The low variance indicates that the framework's performance is a stable property of the models and data rather than a consequence of a favourable partition.

Overall, the results show that academic risk prediction can be performed effectively using lightweight and interpretable machine learning models. The primary dataset demonstrates the feasibility of three-class dropout-oriented risk prediction, while the secondary dataset supports the transferability of the risk-mapping framework to B.Tech semester-grade analysis.

6 Ablation Study and Comparison with Existing Work

6.1 Ablation Study

An ablation study was conducted on the UCI dataset to examine the contribution of different feature groups. The purpose of this analysis was to determine whether academic risk prediction depends mainly on academic progress indicators or whether demographic, socio-economic, and administrative variables contribute equally.

The ablation design separates the feature space into five groups: all features, academic progress features only, administrative/contextual features only, socio-demographic/macro features only, and academic plus administrative features. The resulting macro F1-scores quantify how much predictive signal is retained by each information source (Table 9).

Table 9 Macro F1-score by feature group in the UCI feature-group ablation study

Feature group	Best model	Macro-F1
All features	Logistic Regression	0.6913
Academic + administrative	Logistic Regression	0.6853

Feature group	Best model	Macro-F1
Academic progress only	Logistic Regression	0.6529
Administrative/contextual only	Logistic Regression	0.5451
Socio-demographic/macro only	Logistic Regression	0.4601

The ablation results show that the full feature set provides the highest macro F1-score. However, the academic plus administrative feature group performs very close to the full feature set. This indicates that most of the predictive power comes from academic progress and selected institutional variables.

The ablation plot makes the same pattern visually clear: performance declines from the full feature set to academic plus administrative features, then to academic-only features, and falls sharply when only administrative/contextual or socio-demographic/macro variables are used (Fig. 6).

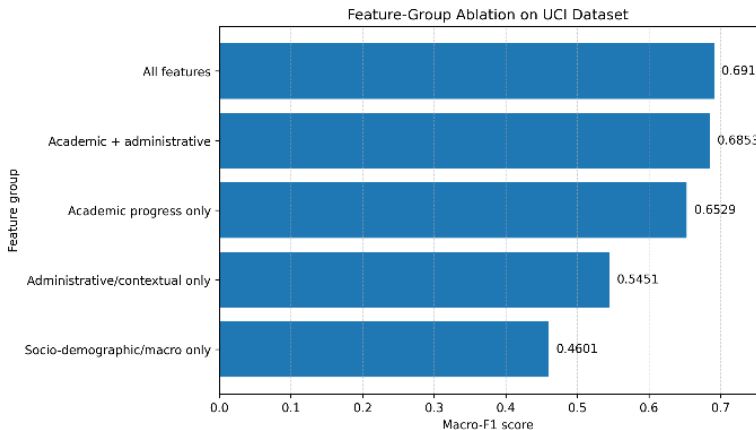


Fig. 6 Macro F1-score comparison of feature groups in the UCI feature-group ablation study

Academic progress features alone achieved a macro F1-score of 0.6529. This is substantially better than the administrative/contextual group and the socio-demographic/macro group. The administrative/contextual group achieved a macro F1-score of 0.5451, while the socio-demographic/macro group achieved only 0.4601. These results clearly show that socio-demographic and macroeconomic variables alone are insufficient for reliable academic-risk prediction.

This finding is important for institutional implementation. It suggests that academic monitoring systems should prioritise semester performance, curricular-unit completion, enrolment progress, and tuition-related status rather than relying heavily on student background attributes. This also has ethical significance. A risk-prediction system that depends mainly on academic progress indicators is more transparent and less likely to overemphasise sensitive demographic factors.

The ablation study strengthens the contribution of this work because it goes beyond model comparison. It identifies which types of data are most valuable for early academic-risk prediction. The results support the claim that academic progression is the strongest practical signal for risk identification, while socio-demographic attributes should be used cautiously and only as supplementary information.

6.2 Comparison with Existing Work

Several recent studies have examined student dropout and academic success prediction using the same or closely related datasets. These studies provide useful benchmarks for positioning the present work.

Martins et al. (2023) proposed a multi-class phased prediction approach for academic performance and dropout in higher education. Their study evaluated prediction at different stages of the first academic year and found that Random Forest-based models performed strongly, especially when imbalance-handling strategies were applied. Their work also showed that prediction improves when academic performance information after enrolment becomes available.

Ridwan et al. (2024) used the UCI dropout and academic success dataset with XGBoost and reported strong performance. However, their work simplified the prediction problem into a binary classification task by grouping student outcomes into success and dropout categories. This makes their reported F1-score less directly comparable with the present study because binary classification is generally easier than three-class risk prediction.

Vo et al. (2026) evaluated a broader set of models, including Logistic Regression, Naive Bayes, k-nearest neighbours, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, XGBoost, LightGBM and CatBoost. Their results showed that Random Forest achieved the highest accuracy of 0.7797, while LightGBM achieved the best macro F1-score of 0.7082. They also reported that early academic progress indicators were among the strongest predictors.

Against related work on the same or closely related datasets, the proposed framework is positioned not only by macro F1-score but also by its three-level risk mapping, dual-dataset validation, explainability and intervention-oriented design (Table 10).

Table 10 Comparison with related work on the same or closely related datasets

Study	Main method	Key result
Martins et al. (2023)	Random Forest with balancing	Best after first semester
Ridwan et al. (2024)	XGBoost	Binary F1 about 0.81
Vo et al. (2026)	Random Forest, LightGBM	Macro-F1 about 0.7082
Proposed work	Lightweight explainable AI	Macro F1-score 0.7112

The proposed framework achieves a macro F1-score of 0.7112 on the primary three-class dataset using a histogram-based gradient-boosting model with engineered academic-progress features, slightly exceeding the best ensemble macro F1-score reported by Vo et al. (0.7082) while remaining lightweight, interpretable, and easy to deploy. Logistic Regression provides a fully transparent linear baseline at 0.6963, only marginally behind.

The main difference between the proposed work and existing studies is not merely the choice of classifier. The present study contributes a unified academic-risk framework that maps different academic prediction problems into High Risk, Moderate Risk, and Low Risk categories. It also validates the framework on a secondary B.Tech dataset, which adds engineering-education relevance. In addition, the study includes a feature-group ablation analysis, which identifies the relative contribution of academic, administrative, and socio-demographic variables. Compared with prior works that focus mainly on predictive performance, the proposed framework gives equal importance to usability and explainability. The model outputs are linked with academic intervention logic. High Risk students can be prioritised for urgent mentoring and remedial support. Moderate Risk students can be placed under close observation and targeted follow-up. Low Risk students can continue under regular academic monitoring.

Therefore, the novelty of the present work lies in the integration of seven elements: three-level academic-risk mapping, lightweight model tuning, dual-dataset validation, class-imbalance-aware evaluation, feature-group ablation, multi-level explainability extended to student-level evidence, and a tiered human-in-the-loop intervention pathway whose robustness is confirmed through threshold-sensitivity and multi-seed reliability analyses. The study does not claim to introduce a new machine learning algorithm. Instead, it provides a practical, reproducible, and institutionally meaningful framework for academic-risk prediction and sustainable student support.

7 Ethical and Practical Considerations

Academic-risk prediction systems should be used as human-in-the-loop decision-support tools, not as automated mechanisms for labelling or penalising students. In the proposed framework, risk predictions and probabilities are intended to prioritise timely mentoring, remedial support and progress monitoring, with faculty mentors reviewing each case in context before action is taken. The framework emphasises academic progression and administrative engagement indicators because they are directly linked to student support, while socio-demographic variables are treated cautiously due to their limited standalone predictive value.

In deployment, institutional ERP, learning-management, attendance and assessment records may be scored periodically, and mentor-facing risk evidence cards can guide Tier 1, Tier 2 or Tier 3 interventions. Student data should be anonymised during model development, prediction access should be restricted to authorised academic personnel, and the system should be communicated as a support mechanism. Future work may extend the framework through multi-institutional longitudinal validation, earlier within-semester prediction and formal fairness auditing.

8 Conclusion and Future Work

This study presented a lightweight explainable artificial intelligence framework for academic-risk prediction using three intervention-oriented risk levels. On the primary dropout dataset, histogram-based gradient boosting achieved a macro F1-score of 0.7112, while Extra Trees achieved a macro F1-score of 0.9708 on the B.Tech dataset. Ablation and explainability analyses showed that academic-progress and administrative variables dominate prediction, whereas socio-demographic variables contribute limited predictive value when used alone. Threshold-sensitivity and multi-seed testing further supported the robustness of the proposed risk-mapping approach.

The framework is best positioned as transparent decision support for mentor-led intervention, not as an automated labelling mechanism. Future work should validate the framework across larger multi-institutional and longitudinal datasets, test earlier within-semester prediction points, integrate attendance and learning-management signals, and conduct a formal fairness audit across relevant student subgroups.

References

- Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., & Al-Fuqaha, A. (2024). Data-driven artificial intelligence in education: A comprehensive review. *IEEE Transactions on Learning Technologies*, 17, 12-31. <https://doi.org/10.1109/TLT.2023.3314610>
- Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. *Education Sciences*, 11(9), Article 552. <https://doi.org/10.3390/educsci11090552>
- Alhazmi, E., & Sheneamer, A. (2023). Early predicting of students performance in higher education. *IEEE Access*, 11, 27579-27589. <https://doi.org/10.1109/ACCESS.2023.3250702>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Butt, N. A., Mahmood, Z., Shakeel, K., Alfarhood, S., Safran, M., & Ashraf, I. (2023). Performance prediction of students in higher education using multi-model ensemble approach. *IEEE Access*, 11, 136091-136108. <https://doi.org/10.1109/ACCESS.2023.3338977>
- Latif, G., Abdelhamid, S. E., Fawagreh, K. S., Brahim, G. B., & Alghazo, R. (2023). Machine learning in higher education: Students' performance assessment considering online activity logs. *IEEE Access*, 11, 69586-69600. <https://doi.org/10.1109/ACCESS.2023.3292117>

- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, 4765-4774.
- Martins, M. V., Baptista, L., Machado, J., & Realinho, V. (2023). Multi-class phased prediction of academic performance and dropout in higher education. *Applied Sciences*, 13(8), Article 4702. <https://doi.org/10.3390/app13084702>
- Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3, Article 100066. <https://doi.org/10.1016/j.caeai.2022.100066>
- Padhy, N., & Panigrahi, R. (2025). Student grade prediction (Version 1) [Data set]. *Mendeley Data*. <https://doi.org/10.17632/6dgkv6kpr2.1>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11), Article 146. <https://doi.org/10.3390/data7110146>
- Realinho, V., Martins, M. V., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success [Data set]. *UCI Machine Learning Repository*. <https://doi.org/10.24432/C5MC89>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). <https://doi.org/10.1145/2939672.2939778>
- Ridwan, A., Priyatno, A. M., & Ningsih, L. (2024). Predict students' dropout and academic success with XGBoost. *Journal of Education and Computer Applications*, 1(2), 1-8.
- Saleem, F., Ullah, Z., Fakieh, B., & Kateb, F. (2021). Intelligent decision support system for predicting students' e-learning performance using ensemble machine learning. *Mathematics*, 9(17), Article 2078. <https://doi.org/10.3390/math9172078>
- Şekeroğlu, B., Abiyev, R., İlhan, A., Arslan, M., & Idoko, J. B. (2021). Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies. *Applied Sciences*, 11(22), Article 10907. <https://doi.org/10.3390/app112210907>
- Vo, N. T., Le, Q. N., & Le, H. (2025). Artificial intelligence-driven early prediction of student dropout and academic outcomes in higher education: A comparative study of advanced machine learning approaches. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, 13(1). <https://doi.org/10.4108/eetinis.131.11758>
- Zafari, M., Sadeghi-Niaraki, A., Choi, S.-M., & Esmaeily, A. (2021). A practical model for the evaluation of high school student performance based on machine learning. *Applied Sciences*, 11(23), Article 11534. <https://doi.org/10.3390/app112311534>