INTERNATIONAL JOURNAL OF RESEARCH, EXPLORATION & KNOWLEDGE (IJREK)
Official Scholarly Journal Portal: **https://www.ijrek.org** | DOI: **https://doi.org/10.5281/zenodo.23482540**

EXPLAINABLE AI FOR EARLY PREDICTION OF STUDENT ACADEMIC RISK USING MACHINE LEARNING

Dr.Bere Sir[1], Srushti Ahire[2]

[1] Faculty Guide & Supervisor | [2] Dattakala Group Of Institutions Faculty Of Engineering Swami - Chincholi tal. Daund dist. Pune, India

Corresponding author: **srushtiahire430@gmail.com**

ARTICLE ID	ARTICLE TYPE	RESEARCH AREA	PUBLICATION
IJREK-2026-00101	Research Article	Artificial Intelligence	Volume 1, Issue Issue 1 (2026)
RECEIVED	**REVISED**	**ACCEPTED**	**PUBLISHED**
19 October 2026	02 November 2026	19 October 2026	02 November 2026

Abstract

Student academic risk prediction aims to identify learners who may need timely academic support before final outcomes are known. Machine learning can analyse historical indicators such as attendance, assessment performance, assignment completion, and learning-platform engagement to estimate the likelihood of academic difficulty. However, a prediction is more useful to educators when its basis can be understood. This paper proposes an explainable machine-learning framework that combines early risk estimation with interpretable explanations for students, faculty, and academic advisors. The workflow includes data governance, preprocessing, model training, validation, risk scoring, local and global explanations, and human review. Candidate models include logistic regression and decision trees as interpretable baselines, alongside more complex models where justified. Explanation methods such as feature importance and SHAP-style contribution analysis can help communicate which recorded indicators influenced a prediction. The system is designed to support—not replace—professional judgement, and it should not automatically penalise students. Fairness, privacy, data quality, and clear communication are treated as core design requirements. The proposed architecture and evaluation plan are presented for an academic prototype; no experimental performance results are claimed.

Keywords: *Explainable AI; student academic risk; machine learning; early prediction; SHAP; educational data mining; fairness; learning analytics*

2. PROBLEM STATEMENT AND OBJECTIVES

The problem addressed is the lack of a transparent, timely, and responsibly governed approach for identifying students who may be at risk of academic difficulty. A useful system should provide an early signal and a comprehensible explanation while allowing staff to review context and correct inaccurate records. The objectives of the proposed work are: • To identify suitable academic indicators for early prediction. • To prepare student records through cleaning, encoding, and missing-value handling. • To compare interpretable baseline models with selected machine-learning models. • To generate explanations showing which features influenced each prediction. • To evaluate predictive performance, calibration, fairness, and explanation usefulness. • To design a dashboard that supports constructive human-led intervention.

3. LITERATURE REVIEW

Educational data mining and learning analytics investigate how data from educational settings can be used to understand learning processes and support student success. Prior studies have applied classification and regression methods to estimate course outcomes, dropout risk, and performance using attendance, prior grades, assessment results, and engagement indicators. Common approaches include logistic regression, decision trees, random forests, support vector machines, and gradient-boosting methods. A major concern is that predictive accuracy alone does not establish that a system is appropriate for educational decisions. Data may be incomplete or biased, target labels may reflect unequal opportunities, and a model trained at one institution may not generalise to another. Explainability methods can help expose model behaviour, but explanations are not proof of causation and may themselves be unstable or misunderstood. This proposed framework therefore treats explainability, privacy, and human oversight as part of the system design rather than optional add-ons. The paper focuses on an implementable workflow and evaluation plan, not on claiming that a specific model has already achieved a particular accuracy.

4. SCOPE AND ASSUMPTIONS

The prototype is intended for early academic support in a defined course or institution. Data fields should be limited to those necessary for the stated purpose, and access should be restricted to authorised personnel. Sensitive personal attributes should not be used casually as predictors; where lawful and ethically approved, carefully governed subgroup analysis may be needed to assess disparities. Risk categories and intervention rules must be agreed with educators before deployment.

5. PROPOSED SYSTEM ARCHITECTURE

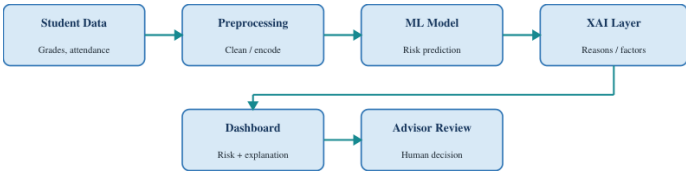


Figure 1. Proposed architecture for explainable student academic risk prediction.

5.1 DATA COLLECTION LAYER

Potential features include attendance percentage, scores from completed assessments, assignment submission patterns, prior course results, and learning-platform activity where its use is appropriate and disclosed. Data should be collected under institutional policy, with access limited by role and retention rules. Records should use a pseudonymous identifier wherever possible.

5.2 PREPROCESSING LAYER

Preprocessing checks duplicate records, inconsistent formats, missing values, outliers, and leakage from future information. Features must only use information available at the prediction date. Categorical variables may be encoded, numeric variables scaled when required, and missingness handled using a documented

method. Training, validation, and test sets should be separated by time or student as appropriate to avoid overly optimistic estimates.

5.3 PREDICTION AND EXPLAINABILITY LAYERS

The prediction layer estimates a risk score or class. The explanation layer describes influential features for the individual prediction and summarises patterns across the model. For example, low recent assessment scores or repeated absences might contribute to a higher estimated risk, but explanations must reflect actual model outputs and should not imply that a feature causes the outcome.

5.4 DASHBOARD AND HUMAN REVIEW

The dashboard presents the risk estimate, explanation, data timestamp, uncertainty or confidence information where available, and links to verified academic records. An advisor reviews the information, checks context, and selects a supportive response. Students should have an appropriate route to correct inaccurate data or request review.

6. METHODOLOGY AND WORKFLOW

1. Collect Approved records 2. Prepare Clean / split 3. Predict Risk probability 4. Explain Feature effects 5. Support Advisor action PROPOSED PREDICTION AND SUPPORT WORKFLOW



Figure 2. Proposed workflow from approved data to human-led support.

6.1 DEFINE THE PREDICTION TARGET

The target must be defined clearly, such as failure to meet a course threshold by a specified date or the need for academic support under an institution-approved rule. It should not be an ambiguous label that changes between departments. The prediction horizon and intended intervention must be defined before modelling.

6.2 MODEL DEVELOPMENT

Begin with interpretable baselines such as logistic regression and a shallow decision tree. More complex models may be compared if there is a clear reason and sufficient data. Use cross-validation within the training set for model selection, and reserve a test set for final evaluation. Hyperparameter tuning must not use the test set. If the data are imbalanced, report class-specific metrics rather than accuracy alone.

6.3 EXPLAINABILITY METHODS

Global feature importance can describe which variables the model relies on overall. Local explanations can describe the factors contributing to a particular prediction. SHAP-style additive attributions are one option for selected model types; simple models may also support direct coefficient or decision-path explanations. Explanations should be checked for stability and presented in plain language, with clear separation between correlation and causation.

6.4 INTERVENTION WORKFLOW

A high-risk signal should initiate a review rather than an automatic penalty. Advisors can verify attendance or assessment records, discuss barriers with the student, and offer relevant academic support. The system should record the reason for an intervention without exposing unnecessary personal information. Risk scores should be periodically reviewed for drift and unintended effects.

6.5 DATA PROTECTION

Use role-based access, secure transport, appropriate encryption, audit logs, and a defined retention schedule. Student records should not be copied to personal devices or public repositories. The institution should explain what data are used, who can access predictions, and how students can raise concerns.

7. RISK VISUALISATION AND EXPLAINABLE OUTPUT

A dashboard should avoid presenting a risk score as a fixed label about a student. The score is an estimate based on available data and may change as new information arrives. Explanations should include the most relevant contributing features, the period covered by the data, and a reminder that staff must verify the underlying records. Illustrative sample index (not actual student data) Risk score (%) EXAMPLE VISUALISATION OF MODEL RISK SCORES



Figure 3. Illustrative risk-score trend for interface design only; it is not actual student data or a model result.

7.1 EXAMPLE OF AN EXPLANATION

A hypothetical explanation might state: “The estimated risk increased mainly because recent assessment scores declined and several assignments were submitted late.” This wording is appropriate only if those features genuinely influenced the trained model for that prediction. The interface should also show positive factors where relevant and allow the advisor to inspect the underlying data.

7.2 DASHBOARD COMPONENTS

• Risk estimate and prediction date. • Main contributing factors with plain-language descriptions. • Attendance and assessment trends, with links to source records. • Data-quality flags for missing or outdated information. • Advisor review status and supportive actions offered. • A correction or appeal route for students and authorised staff.

7.3 COMMUNICATION PRINCIPLES

Use neutral wording such as “may benefit from support” rather than labelling a learner as weak or unsuccessful. Avoid implying certainty, avoid ranking students publicly, and do not display explanations to people without a legitimate educational role. Model outputs should be contextualised alongside teacher knowledge and student voice.

8. IMPLEMENTATION PLAN

Table 1. Proposed system modules

Module	Responsibility	Important control
Data ingestion	Import approved academic records	Access permissions and audit trail
Preprocessing	Clean, encode, and validate features	Prevent leakage; document missing data
ML model	Estimate academic risk	Validation and calibration
XAI layer	Explain influential features	Avoid causal overclaiming
Dashboard	Show score, trend, and explanations	Role-based visibility
Advisor review	Verify context and offer support	No automatic punitive decisions

8.1 SUGGESTED TECHNOLOGY STACK

Python can be used for data preparation and model development with common machine-learning libraries. A lightweight API can expose approved predictions to a web dashboard, while a relational database stores authorised records and audit events. The exact technology choices should follow institutional security requirements and the available project environment.

9. EVALUATION FRAMEWORK

TABLE II

Evaluation area	Measure or method	Reason
Classification quality	Precision, recall, F1-score; ROC-AUC where suitable	Measure errors and detection ability
Calibration	Compare predicted probabilities with observed outcomes	Assess score reliability
Fairness	Compare errors and outcomes across approved groups	Identify potential disparities
Explanation quality	Advisor review, usefulness, and stability checks	Ensure reasons are understandable
Operational utility	Time to review, data correction, support follow-up	Assess real-world usefulness

11. LIMITATIONS AND ETHICAL CONSIDERATIONS

Predictions can be inaccurate when data are sparse, outdated, or not representative. Attendance and assessment patterns may reflect health, disability, employment, transport, family responsibilities, or other circumstances that the dataset does not capture. A model may reproduce historic inequities or create false alarms. Explainability does not remove these risks. The institution should define accountability, provide human review, audit performance over time, restrict access, and allow correction of erroneous records. Predictions must not be used as the sole basis for punitive or high-stakes decisions.

12. EXPECTED OUTCOMES

The proposed outcome is a prototype that produces an early academic-risk estimate, presents understandable feature contributions, and supports advisor review. The effectiveness of the approach remains to be established through a properly governed dataset, rigorous testing, and feedback from intended users.

13. FUTURE SCOPE

Future work may examine temporal models, course-specific prediction, model monitoring for changing student populations, and explanation interfaces tested with students and advisors. Additional work can investigate privacy-preserving learning and methods for assessing explanation stability. Any extension should be introduced only after confirming its educational value, fairness, and compliance with institutional policy. 14. Conclusion This paper proposes an explainable machine-learning framework for early prediction of student academic risk. It combines approved academic indicators, preprocessing, predictive models, local and global explanations, and a dashboard for human review. The approach aims to make early-warning signals more understandable and actionable while acknowledging that risk estimates are uncertain and may reflect limitations in the data. Responsible use requires privacy protection, fairness evaluation, clear communication, student data-correction mechanisms, and supportive human-led interventions. The proposed system should be validated empirically before being used in routine educational decisions.

REFERENCES

[1] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, no. 6, pp. 601–618, 2010.

[2] R. S. J. d. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[3] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020.

[4] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, 2017.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘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, 2016.

[6] UNESCO, Guidance for Generative AI in Education and Research. Paris: UNESCO, 2023.

[7] OECD, OECD Principles on Artificial Intelligence. Organisation for Economic Co-operation and Development, 2019.

[8] U.S. Department of Education, Office of Educational Technology, Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations, 2023.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on **https://www.ijrek.org** | Article ID: IJREK-2026-00101