INTERNATIONAL JOURNAL OF RESEARCH, EXPLORATION & KNOWLEDGE (IJREK)
Official Scholarly Journal Portal: **https://www.ijrek.org** | DOI: **https://doi.org/10.5281/zenodo.23862413**

EXPLAINABLE MACHINE LEARNING-BASED STUDENT ACADEMIC PERFORMANCE PREDICTION

Ms. Gawand R. R.[1], Ritesh Ovhal[2]

[1] Faculty Guide & Supervisor | [2] Dattakala Group of Institutions, India

Corresponding author: **ovhalritesh500@gmail.com**

ARTICLE ID	ARTICLE TYPE	RESEARCH AREA	PUBLICATION
IJREK-2026-00179	Research Article	Computer Science	Volume 1, Issue Issue 1 (2026)
RECEIVED	**REVISED**	**ACCEPTED**	**PUBLISHED**
29 October 2026	02 November 2026	30 October 2026	03 November 2026

Abstract

Student academic performance prediction can help educational institutions identify learners who may benefit from timely, targeted support. Machine-learning models can combine carefully selected indicators such as prior assessment results, attendance, assignment completion and learning-platform engagement to estimate likely performance. However, a prediction that cannot be explained may be difficult for teachers and students to trust, and may conceal data quality problems or unfair patterns. This paper presents a conceptual framework for explainable machine-learning-based student academic performance prediction. The proposed workflow includes data governance, preprocessing, feature engineering, baseline and advanced model comparison, interpretable explanations, fairness checks, and human-led intervention. Candidate models include logistic regression, decision trees, random forests and gradient-boosted trees, with SHAP- or permutation-based explanations where appropriate. Evaluation should use held-out data, class-sensitive metrics, calibration and subgroup analysis rather than accuracy alone. The framework is designed to support—not replace—professional educational judgement. No original dataset or experimental result is claimed; model performance must be established through a documented local study.

Keywords: *student performance; explainable AI; machine learning; educational data mining; early warning; fairness; SHAP*

2. BACKGROUND AND RELATED WORK

Educational data mining and learning analytics study data generated during teaching and learning to understand progress and improve support. Prior work commonly explores classification of pass/fail or risk categories, regression of final scores, and time-sensitive prediction from activity sequences. Simple models can provide useful baselines, while ensemble methods may capture nonlinear interactions. Their relative value depends on dataset size, missingness, target definition and how the model will be used. Explainable artificial intelligence (XAI) covers methods that make model behaviour more understandable. Intrinsically interpretable models— such as linear models and small decision trees—can expose their logic directly. Post-hoc methods, including feature permutation and SHAPstyle attribution, approximate how features contribute to a model output. Explanations are not causal proof: a feature may correlate with an outcome because it proxies for another variable, reflects institutional practice, or encodes historical inequity. 2.1 Limitations in common approaches A high overall accuracy can hide poor recall for students who need support, and a random split can overstate performance when records from the same student or semester appear in both training and test sets. Leakage may occur when a feature is recorded after the outcome being predicted, such as a final attendance total used to predict an earlier warning. Models can also become stale after changes in curriculum, grading or online-learning platforms. These issues require explicit study design, not just algorithm selection. 2.2 Research gap and contribution This paper organises prediction, explanation, fairness, privacy and intervention into one practical framework. The emphasis is on traceability: each prediction should have a defined target and time horizon, documented inputs, a confidence or uncertainty signal, an explanation, and a route for human review. The framework is adaptable to different institutions and does not prescribe a single universally best model.

3. PROPOSED SYSTEM OVERVIEW

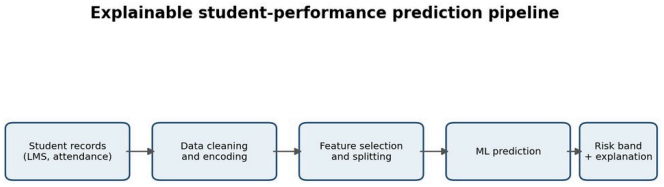


Figure 1. End-to-end workflow from authorised student records to an interpretable prediction.

The proposed system has six logical components: data ingestion; validation and preprocessing; feature construction; model training and evaluation; explanation and risk communication; and educator review with outcome monitoring. Access controls and audit logs apply across all components. Data should be minimised to what is necessary for the stated educational purpose, and students should receive clear information about use wherever required by policy and applicable law.

4. DATA SOURCES AND FEATURE ENGINEERING

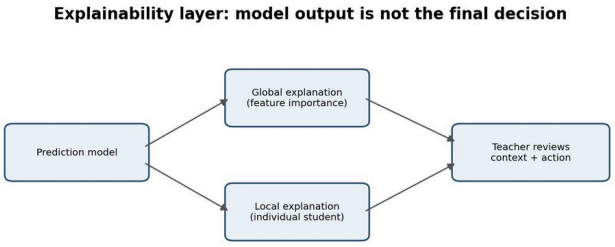
TABLE I

Data group	Possible features	Key caution
Assessment	Previous marks, quiz trend, submission scores	Use only marks available at prediction time
Engagement	LMS logins, resource views, practice activity	Activity volume is not equivalent to learning
Attendance	Attendance to date, recent change	Check legitimate absences and recording errors
Course context	Course load, assessment schedule, prerequisites	Avoid comparing unlike courses without adjustment
Support history	Advising contact or tutoring participation	Sensitive; may reflect prior institutional decisions

5. MACHINE-LEARNING METHODOLOGY

TABLE II

Model	Strengths	Limitations / use
Logistic / linear regression	Transparent baseline; coefficients can be inspected	May miss complex nonlinear relationships
Decision tree	Readable if shallow; rule-based explanations	Can be unstable and overfit without constraints
Random forest	Handles interactions and mixed patterns	Individual decisions are less transparent
Gradient-boosted trees	Often strong on tabular data	Needs tuning and careful explanation
Neural network	Flexible for large or sequential data	Requires more data and stronger explanation controls



5.3 System architecture

Figure 2. Explanation is provided to the educator alongside the model output; it does not independently determine action. 5.4 Reproducible implementation stores the data schema, preprocessing pipeline, model version, evaluation report and explanation settings. A prediction record should identify the model version and time of scoring. Re-training should be controlled, validated and approved rather than silently replacing the deployed model.

6. EXPLAINABILITY AND COMMUNICATION

Explainability is central when predictions may influence educational support. Global explanations describe patterns used across many predictions, such as the relative contribution of recent assessment results or assignment completion. Local explanations describe factors associated with one individual prediction. Both should be accompanied by caveats: explanations depend on the model, the available features and their correlations, and may vary when inputs are similar or unstable. 6.1 Explanation methods For linear models, coefficients can indicate direction and magnitude after accounting for feature scaling and encoding. Small decision trees can show explicit decision paths. Permutation importance estimates the loss in model quality when a feature is disrupted, though correlated features can share importance. SHAP-style methods allocate a model output among features under a defined reference distribution. These methods answer different questions and should not be treated as interchangeable or causal. 6.2 Student-centred explanation An explanation should use plain language and focus on changeable learning conditions. For example, a system might indicate that recent quiz performance and several unsubmitted assignments contributed to a risk estimate. It should avoid statements such as “the student is lazy” or “the student will fail.” Where possible, show uncertainty, data freshness, the date of the prediction and a route to correct inaccurate records. The interface should not imply that a low-risk prediction guarantees success. 6.3 Human-in-the-loop process

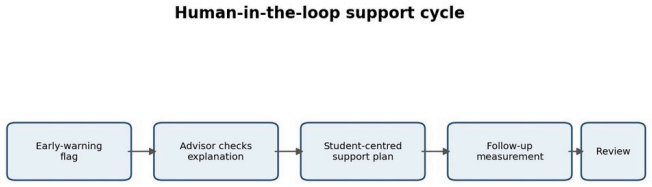


Figure 3. A prediction initiates review and supportive action; subsequent outcomes inform system evaluation.

7. EVALUATION FRAMEWORK

TABLE III

Metric	Meaning	Interpretation note
Accuracy	Correct predictions / all predictions	Can mislead with imbalanced outcomes
Precision	True positives / predicted positives	How many alerts are relevant
Recall / sensitivity	True positives / actual positives	How many students needing support are found
F1-score	Harmonic mean of precision and recall	Useful summary, but hides trade-offs
ROC-AUC / PR-AUC	Ranking quality across thresholds	PR-AUC often informative for rare positives
MAE / RMSE	Average absolute / squared-error-based score error	Report in meaningful score units
Calibration	Agreement of predicted and observed risk	Important if probabilities are shown

7.2 ROBUSTNESS AND DRIFT

Test the model on later semesters, different courses and plausible data-quality disruptions. Monitor changes in feature distributions, missingness, calibration and intervention rates. Curriculum changes, new assessment formats and altered LMS logging can cause drift. Define thresholds that trigger investigation, revalidation or temporary suspension. Maintain a rollback plan if model quality deteriorates. 7.3 Illustrative evaluation plan A local study could compare a majority-class or mean baseline with logistic/linear regression, a shallow tree and an ensemble. Pre-register the target, split strategy and primary metric; tune only on training data; and publish the test-set results with limitations. Because no institutional dataset was supplied for this paper, no numerical results or claim of a best-performing model is presented.

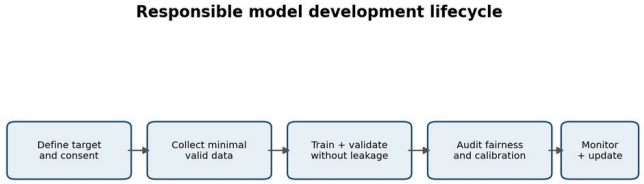


Figure 4. Validation and monitoring should continue throughout the model lifecycle.

8. PRIVACY, ETHICS AND DEPLOYMENT

Student records can reveal personal circumstances and should be handled as sensitive information. Institutions should establish a clear purpose, lawful basis, access roles, retention period and process for correction. Use data minimisation, encryption in transit and at rest, secure backups and audit logs. Identifiers should be removed or

pseudonymised for analysis where practical, recognising that re-identification can still be possible when records are linked.

8.1 GOVERNANCE AND ACCOUNTABILITY

A designated institutional owner should be responsible for approving use, reviewing impact and responding to complaints. Educators should be trained on model limitations and explanation displays. Students should be able to challenge incorrect records and seek human review. The system should not use risk estimates to stigmatise, publicly rank or reduce opportunities for students. Any high-impact decision requires independent human judgement and applicable institutional/legal safeguards. 8.2 Deployment architecture and controls A practical deployment can separate the student-information source, a controlled data-processing service, a model API and an educator dashboard. The dashboard should present prediction time, model version, uncertainty, key contributing factors and data-quality warnings. Rolebased access should limit individual-level views to staff with a legitimate need. Logs should capture access and overrides without unnecessarily duplicating sensitive data.

8.3 RISKS AND MITIGATIONS

TABLE IV

Risk	Mitigation
Incorrect or stale records	Show source/time; provide correction workflow
False positive alerts	Human review; thresholds tuned with staff; avoid punitive use
Missed at-risk students	Monitor recall; retain non-model support routes
Proxy discrimination	Feature review, subgroup audits and documented mitigation
Data breach	Least privilege, encryption, retention limits and incident plan
Over-reliance on prediction	Training, uncertainty display and override rights

9. DISCUSSION, LIMITATIONS AND FUTURE SCOPE

The proposed framework positions explainability as a component of a responsible educational decision-support process rather than an add-on. The main value is the combination of prediction, transparent evidence, human review and follow-up measurement. A technically accurate model is not automatically educationally useful: the institution must have appropriate support services and evaluate whether alerts lead to beneficial outcomes. 9.1 Limitations This work is a conceptual review and system proposal, not an empirical evaluation. No student dataset, institutional approval, model training run or intervention study was provided. Consequently, it cannot establish accuracy, fairness, generalisability or impact on retention. Feature relevance varies across institutions, courses and age groups. Post-hoc explanations have known limitations, and a feature contribution should not be interpreted as proof that changing the feature will change the outcome. 9.2 Future scope Future work can examine temporal models that use the sequence of learning events, calibrated uncertainty estimates, counterfactual explanations that suggest feasible next steps, and privacy-preserving analysis across institutions. Any use of generative AI should be evaluated separately and should not expose identifiable student records to unapproved services. Prospective studies should assess outcomes such as timely support, student experience and academic progress, while considering unintended effects such as alert fatigue or stigma. 10. Conclusion Explainable machine learning can support earlier identification of students who may benefit from academic assistance when it is developed with sound data practices and clear educational goals. A robust approach defines the target and prediction time, prevents leakage, compares transparent baselines with more complex models, evaluates calibration and subgroup errors, and communicates predictions in plain language. Privacy, fairness, uncertainty and human oversight must be built into the lifecycle. Most importantly, predictions should trigger supportive conversations and access to resources—not deterministic labels or punitive decisions. A local, ethically governed validation study is required before deployment, and ongoing monitoring is necessary as courses and student populations change.

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, vol. 40, no. 6, pp. 601–618, 2010.

[2] G. Siemens and R. S. J. d Baker, “Learning analytics and educational data mining: Towards communication and collaboration,” in Proc. 2nd Int. Conf. Learning Analytics and Knowledge, 2012, pp. 252–254.

[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, vol. 30, 2017.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’ Explaining the predictions of any classifier,” in Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[6] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] A. P. D. B. et al., “Learning analytics: From data to impact,” (general background on learning analytics; verify exact bibliographic details before formal submission).

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

[1]–[6] and

[8] are established background sources. Verify the exact citation details and add institution-specific or recent peer-reviewed studies before formal publication; reference

[7] is intentionally flagged for replacement rather than presented as a verified citation.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on **https://www.ijrek.org** | Article ID: IJREK-2026-00179