

Predicting Student Academic Success with Machine Learning: An Early-Warning Study for AI-Driven Personalized Learning

SHAMAL KOLHE

shamal.kolhe_24uds@sanjivani.edu.in

sanjivani university <https://orcid.org/0009-0008-5723-1260>

kartik phopase

sanjivani university <https://orcid.org/0009-0002-4168-179X>

Research Article

Keywords:

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11402550/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Predicting Student Academic Success with Machine Learning: An Early-Warning Study for AI-Driven Personalized Learning

Shamal Kolhe

ORCID:  0009-0008-5723-1260

Department of Artificial Intelligence and Data Science
Sanjivani University, Maharashtra, India

Preprint, October 2026

Abstract—Personalized learning systems need a reliable way to identify students who are likely to struggle so that adaptive support can be delivered early. This paper studies that problem as a binary classification task on the public UCI Student Performance (mathematics) dataset of 395 secondary-school students, where the target is whether a student passes the final exam (grade ≥ 10 of 20). Four machine learning models (logistic regression, RBF support vector machine, random forest, gradient boosting) and a majority-class baseline are compared with repeated stratified 5×10 -fold cross-validation on a 316-student training split and a single evaluation on a 79-student held-out test split. Two scenarios are examined: (A) an early-warning scenario that uses only demographic, family, and behavioural attributes, and (B) a mid-term scenario that additionally uses the first- and second-period grades. In scenario A, models reach a cross-validated ROC-AUC of only 0.66–0.69, and the selected model does not beat the majority-class baseline on the held-out set (accuracy 0.658 vs. 0.671; AUC 0.590, 95% CI 0.451–0.723). In scenario B, gradient boosting achieves a held-out accuracy of 0.899, F1 of 0.920 and AUC of 0.945 (95% CI 0.889–0.987), and the second-period grade dominates the permutation importance. The results show that prior-grade information is what makes accurate prediction possible, and that background attributes alone are weak predictors. These findings are relevant for designing the learning-analytics component of AI agent-based adaptive learning systems.

Index Terms—Personalized Learning, Machine Learning, Student Performance Prediction, Educational Data Mining, Learning Analytics, Early Warning Systems

I. INTRODUCTION

Artificial Intelligence (AI) is increasingly used in education for intelligent tutoring, automated assessment, adaptive learning, and personalized learning experiences. Traditional teaching generally gives all students the same materials and explanations, although students differ in ability, prior knowledge, pace, and preferences. Personalized learning systems therefore need a dependable way to detect which students are likely to struggle so that support can be adapted early.

Large Language Models can generate explanations, answer questions, create learning material, and provide feedback [1–3]. Such tutoring components work best when paired with a learning-analytics component that estimates each learner’s risk and knowledge state. Machine learning classifiers are a natural basis for that component, but how well they perform depends strongly on which student data are available at the time of prediction.

This paper therefore asks a focused empirical question: *how accurately can standard machine learning models predict whether a student will pass, with and without access to earlier grades?* The main contributions are:

- A reproducible comparison of four machine learning models and a baseline on a public student-performance dataset.
- Two prediction scenarios—early warning (no grades) and mid-term (with grades)—evaluated under one protocol.
- An honest account of where prediction works (with prior grades) and where it does not (background attributes alone),

with uncertainty estimates and permutation-based feature importance.

II. PROBLEM STATEMENT

A. Lack of Personalization in Traditional Systems

A one-size-fits-all approach does not consider individual learning abilities, knowledge levels, or pace, which can reduce engagement and academic performance. Some students need additional support on fundamentals while others are ready for advanced topics.

B. Need for Reliable Early Identification of At-Risk Students

AI-based learning platforms often provide automated assistance and basic recommendations, but adaptive decisions are only as good as the underlying estimate of student risk. It is unclear how much of that estimate can be obtained from background and behavioural data alone, before any assessment result exists, and how much depends on prior grades.

III. RESEARCH OBJECTIVES

1. To develop and compare machine learning models that predict whether a student passes the final exam.
2. To evaluate prediction quality in an early-warning scenario (no prior grades) and a mid-term scenario (with prior grades), and to identify the most influential attributes.

IV. LITERATURE REVIEW

A. Large Language Models in Personalized Learning

Sharma et al. reviewed the role of Large Language Models in personalized learning and found that they can improve engagement, provide adaptive experiences, and enable real-time feedback, while raising privacy, ethics, and responsible-adoption challenges [1]. Guizani et al. examined LLMs in higher education and reported support for intelligent tutoring, automated assessment, and personalization, with concerns about fairness, privacy, AI dependency, and reliability [2]. Xu et al. [3] and Gan et al. [4] surveyed LLM applications such as tutoring, content generation, question answering, and assessment support.

B. Knowledge Tracing and AI-Based Personalized Learning

Cho et al. reviewed knowledge tracing together with LLMs in education [5]. Farhood et al. found that AI can improve personalization, engagement, and adaptive content delivery [6]. Hariyanto et al. examined techniques used in adaptive education, including machine learning, deep learning, and reinforcement learning [7], and Merino-Campos reported improvements in engagement, performance, and efficiency in higher education [8]. Further reviews cover global advances in AI-based personalized learning [9] and generative AI for teaching and learning [10].

C. Predicting Secondary-School Performance

Cortez and Silva applied data mining to predict the performance of Portuguese secondary-school students and released the dataset used in this study [11]. Their work showed that prior-period grades are highly informative, which motivates the two scenarios examined here.

V. RESEARCH GAP

The literature shows that AI, adaptive learning, LLMs, and learning analytics can improve educational experiences, but many studies treat them independently and report accuracy without clarifying what data were available at prediction time. This study addresses the narrower gap of quantifying how model performance depends on the availability of prior grades, under a single, leakage-free evaluation protocol with uncertainty estimates.

VI. RESEARCH METHODOLOGY

Fig. 1 summarises the pipeline.

A. Dataset

The data are the mathematics-course file of the UCI Student Performance dataset [11]: 395 students from two Portuguese secondary schools, with 33 attributes and no missing values. The attributes comprise 17 nominal (e.g., school, sex, parental job, school/family support), 10 ordinal items (parental education, travel time, study time, family relations, free time, going out, weekday and weekend alcohol use, health; coded on short ordered scales), and 3 ratio-scale counts (age, past failures, absences), plus the grades G1, G2 (first and second period) and

TABLE I. INPUT VARIABLES BY MEASUREMENT SCALE

Scale	Count	Examples
Nominal	17	school, sex, Mjob, Fjob, reason
Ordinal	10	Medu, Fedu, studytime, goout, health
Ratio	3	age, failures, absences
Grades (scenario B only)	2	G1, G2

G3 (final), each on a 0–20 scale. Table I gives the breakdown.

B. Target and Scenarios

The target is $pass = 1$ if $G3 \geq 10$ and 0 otherwise; 67.1% of students pass, so the majority-class baseline has accuracy 0.671. **Scenario A** (early warning) uses the 30 non-grade attributes. **Scenario B** (mid-term) adds G1 and G2 for 32 features. G3 is never used as an input.

C. Preprocessing and Models

Nominal variables are one-hot encoded and numeric variables standardised; both steps are fitted inside each training fold through a single pipeline to avoid leakage. The models are logistic regression (L2, $C=1$), an RBF-kernel SVM ($C=1$), a random forest (500 trees) [12], and gradient boosting (100 trees, depth 3, learning rate 0.1) [13], plus a majority-class baseline. All use scikit-learn 1.8.0 [14] with default settings otherwise; no hyperparameter tuning was performed. The random seed is 42.

D. Evaluation Protocol

Data are split once into a stratified training set (316 students) and a held-out test set (79 students). Models are compared on the training set only, using repeated stratified 5-fold cross-validation with 10 repeats (50 folds). The model with the highest mean cross-validated ROC-AUC in each scenario is refit on the full training set and evaluated *once* on the test set. We report accuracy, precision, recall, F1 (pass is the positive class), and ROC-AUC, with 95% bootstrap confidence intervals (2,000 resamples) on the test set. Feature influence is measured by permutation importance on the test set (30 repeats, drop in ROC-AUC).

VII. RESULTS

A. Cross-Validated Model Comparison

Table II and Fig. 2 show the training-set cross-validation results.

In scenario A, no model clearly separates passing from failing students: AUC values of 0.66–0.69 are modest, and the SVM and random forest reach slightly higher accuracy than the baseline mainly by predicting “pass” for most students (recall 0.92–0.95). Differences between models are smaller than the fold-to-fold standard deviations. In scenario B, all models are strong (AUC 0.95–0.98). Random forest and gradient boosting are tied at 0.977 to three decimals; gradient boosting is selected by a margin below the third decimal, so the choice between them is not meaningful.



Fig. 1. Experimental pipeline used in this study.

TABLE II. CROSS-VALIDATION RESULTS ON THE TRAINING SET (MEAN $\pm$ SD OVER 50 FOLDS; PASS = POSITIVE CLASS)

Scenario	Model	Accuracy	F1	Recall	ROC-AUC
A: no grades	Majority baseline	0.671 $\pm$ 0.006	0.803 $\pm$ 0.004	1.000	0.500
	Logistic Regression	0.689 $\pm$ 0.046	0.784 $\pm$ 0.034	0.843 $\pm$ 0.059	0.693 $\pm$ 0.063
	SVM (RBF)	0.721 $\pm$ 0.035	0.821 $\pm$ 0.021	0.952 $\pm$ 0.034	0.676 $\pm$ 0.060
	Random Forest	0.710 $\pm$ 0.037	0.810 $\pm$ 0.026	0.920 $\pm$ 0.052	0.691 $\pm$ 0.053
	Gradient Boosting	0.688 $\pm$ 0.048	0.782 $\pm$ 0.037	0.837 $\pm$ 0.066	0.661 $\pm$ 0.065
B: with G1, G2	Majority baseline	0.671 $\pm$ 0.006	0.803 $\pm$ 0.004	1.000	0.500
	Logistic Regression	0.921 $\pm$ 0.036	0.941 $\pm$ 0.026	0.945 $\pm$ 0.035	0.975 $\pm$ 0.014
	SVM (RBF)	0.870 $\pm$ 0.039	0.906 $\pm$ 0.029	0.935 $\pm$ 0.042	0.952 $\pm$ 0.022
	Random Forest	0.927 $\pm$ 0.031	0.945 $\pm$ 0.024	0.941 $\pm$ 0.037	0.977 $\pm$ 0.016
	Gradient Boosting	0.922 $\pm$ 0.031	0.942 $\pm$ 0.023	0.943 $\pm$ 0.038	0.977 $\pm$ 0.015

TABLE III. HELD-OUT TEST PERFORMANCE ($n=79$)

Metric	A: Logistic Reg.	B: Grad. Boosting
Accuracy	0.658	0.899
95% CI	0.557–0.759	0.823–0.962
Precision	0.717	0.979
Recall	0.811	0.868
F1	0.761	0.920
ROC-AUC	0.590	0.945
95% CI	0.451–0.723	0.889–0.987
Baseline accuracy	0.671	0.671

B. Held-Out Test Performance

Table III gives the single test-set evaluation of the selected model in each scenario.

In scenario A the selected logistic regression model has a test accuracy of 0.658, *below* the majority baseline of 0.671, and the AUC interval (0.451–0.723) includes chance level (0.5). We therefore conclude that, in this sample, the background and behavioural attributes did not yield a usable predictor. The cross-validated AUC of 0.693 for this model was higher than its test AUC of 0.590, which is consistent with the small test set (79 students) and with mild optimism from model selection.

In scenario B the selected model achieves accuracy 0.899,

F1 0.920, and AUC 0.945, far above the baseline, with a lower interval bound for AUC of 0.889. Fig. 3 and Fig. 4 show the ROC curves and confusion matrices. In scenario B the model misclassified one failing student as passing (1 of 26 failing students) and seven passing students as at-risk (7 of 53). For an early-warning use case, missing a failing student is usually the costlier error, so this error profile is favourable, although the counts are small.

C. Feature Importance

Fig. 5 shows permutation importance. In scenario B the second-period grade G2 dominates (mean AUC drop 0.357), followed at a distance by G1 (0.016); all remaining attributes contribute less than 0.01. In scenario A the largest drops are for past failures (0.042), going out with friends (0.037), age (0.027), and mother’s education (0.019), but the standard deviations (e.g., 0.038 for failures) are of similar size, so the ranking is unstable and should not be over-interpreted.

VIII. DISCUSSION

The central finding is that prior-grade information, not background data, carries the predictive signal. Accuracy rises from the baseline level to about 0.90 once G1 and G2 are available, and G2 alone accounts for most of the importance. This agrees with the original study of this dataset [11] and has a practical implication for personalized learning systems: a

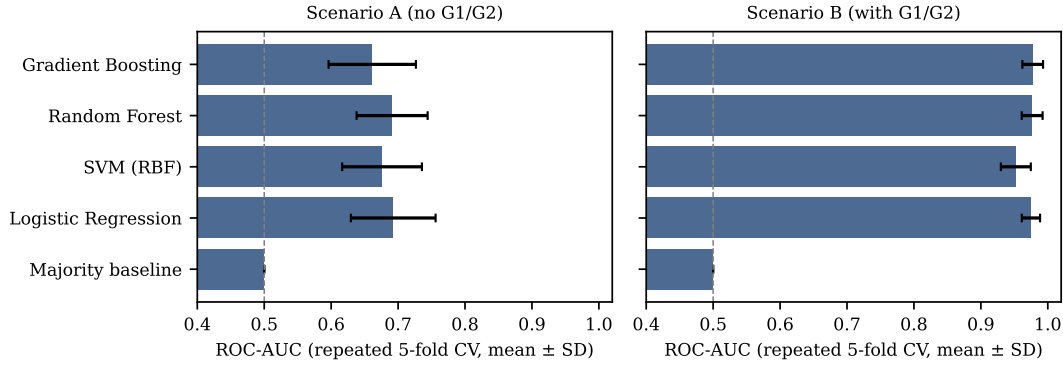


Fig. 2. Cross-validated ROC-AUC by model and scenario. The dashed line marks chance level.

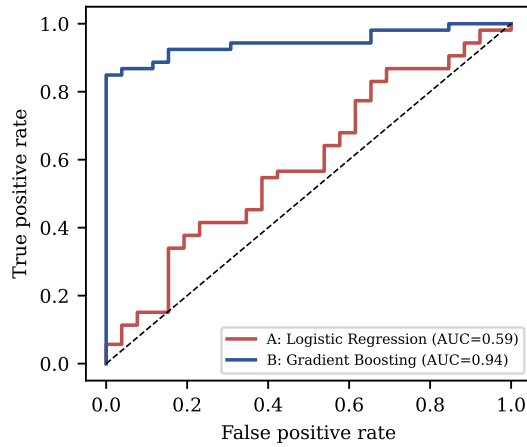


Fig. 3. Test-set ROC curves for the selected model in each scenario.

learning-analytics module should combine continuous assessment data (quiz scores, assignment marks, progress) with any demographic or behavioural data, rather than relying on the latter.

Two cautions apply. First, scenario B is a “mid-term” predictor: because G2 is strongly correlated with G3, it is useful for triggering interventions late in the course, not for very early warning. Second, scenario A suggests that very early, assessment-free prediction on this data is not yet reliable, so adaptive systems should avoid making high-stakes decisions from such attributes. Including attributes such as sex or parental background in a deployed model would also raise fairness concerns that this study did not evaluate.

IX. THREATS TO VALIDITY

Table IV lists the main limitations. In addition, the pass threshold of 10 is a modelling choice; other thresholds or regression of the raw G3 grade could give different conclusions.

		A		B		
Actual	Fail	9	17	Fail	25	1
	Pass	10	43	Pass	7	46
		Fail	Pass	Fail	Pass	
		Predicted		Predicted		

Fig. 4. Test-set confusion matrices (A and B).

TABLE IV. THREATS TO VALIDITY

Threat	Explanation
Small sample	Only 395 students (79 in the test set); intervals are wide and one split was used.
Single context	One course, two schools, one country; results may not transfer.
Untuned models	Default hyperparameters were used; tuning might improve scenario A slightly.
Self-reported items	Survey attributes (alcohol use, free time) may be inaccurate.
Grade dependence	G1 and G2 are strongly tied to G3, which inflates scenario B for early use.
Fairness	Attributes such as sex and family background were not audited for bias.

X. CONCLUSION

This paper compared four machine learning models and a baseline for predicting student pass/fail on the UCI Student Performance (mathematics) dataset. With only background and behavioural attributes, no model reliably beat the majority-class baseline (held-out accuracy 0.658 vs. 0.671; AUC 0.590). With first- and second-period grades, gradient boosting reached a held-out accuracy of 0.899, F1 of 0.920, and AUC of 0.945, with the second-period grade as the dominant predictor. For AI agent-based personalized learning systems, these results indicate that the learning-analytics compo-

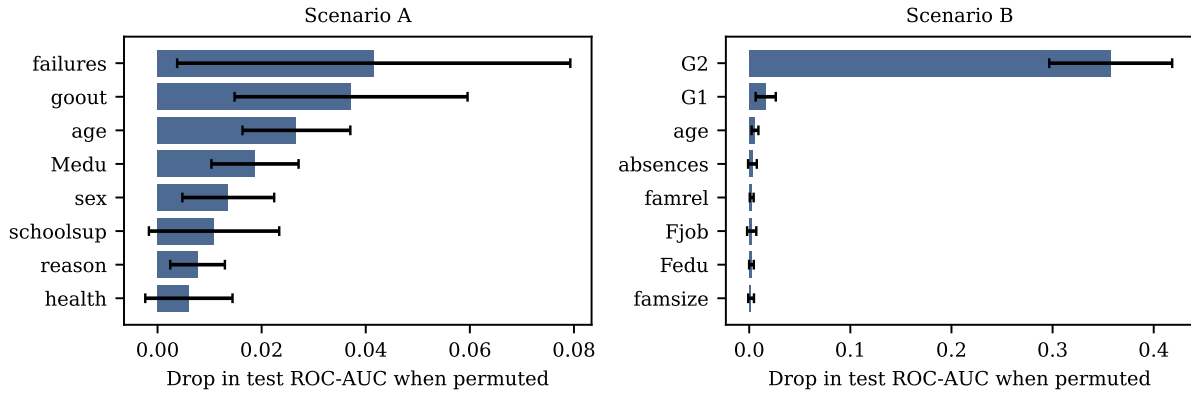


Fig. 5. Permutation importance (test set) for the selected model in each scenario; top eight features, mean $\pm$ SD over 30 repeats.

ment should be driven by assessment data collected during the course.

Future work will use larger and multi-course datasets, repeated outer splits or nested cross-validation, hyperparameter tuning, fairness audits, and integration of the risk predictor with LLM-based explanation and adaptive-quiz modules, followed by evaluation with real students.

DATA AND CODE AVAILABILITY

The dataset is publicly available from the UCI Machine Learning Repository (Student Performance) [11]. All experiments use random seed 42 and scikit-learn 1.8.0.

ACKNOWLEDGMENT

The author would like to thank Sanjivani University and the faculty members for their guidance and support during this research.

REFERENCES

- [1] S. Sharma, P. Mittal, M. Kumar, and V. Bhardwaj, "The role of large language models in personalized learning: a systematic review of educational impact," *Discover Sustainability*, vol. 6, no. 1, p. 243, Apr. 2025, doi: 10.1007/s43621-025-01094-z.
- [2] S. Guizani, T. Mazhar, T. Shahzad, W. Ahmad, A. Bibi, and H. Hamam, "A systematic literature review to implement large language model in higher education: issues and solutions," *Discover Education*, vol. 4, no. 1, p. 35, Feb. 2025, doi: 10.1007/s44217-025-00424-7.
- [3] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu, "Large Language Models for Education: A Survey," arXiv, 2024, doi: 10.48550/ARXIV.2405.13001.
- [4] W. Gan, Z. Qi, J. Wu, and J. C.-W. Lin, "Large Language Models in Education: Vision and Opportunities," in *2023 IEEE Int. Conf. on Big Data*, 2023, pp. 4776–4785.
- [5] Y. Cho, R. E. AlMamlook, and T. Gharaibeh, "A Systematic Review of Knowledge Tracing and Large Language Models in Education: Opportunities, Issues, and Future Research," arXiv, 2024.
- [6] H. Farhood, M. Nyden, A. Beheshti, and S. Muller, "Artificial intelligence-based personalised learning in education: a systematic literature review," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 331, 2025.
- [7] Hariyanto, F. X. D. Kristianingsih, and R. Maharani, "Artificial intelligence in adaptive education: a systematic review of techniques for personalized learning," *Discover Education*, vol. 4, no. 1, p. 458, 2025.
- [8] C. Merino-Campos, "The Impact of Artificial Intelligence on Personalized Learning in Higher Education: A Systematic Review," *Trends in Higher Education*, vol. 4, no. 2, p. 17, 2025.
- [9] A. Fortuna *et al.*, "Artificial intelligence in personalized learning: A global systematic review of current advancements and shaping future opportunities," *Social Sciences and Humanities Open*, vol. 12, p. 102114, 2025.
- [10] B. Ogunleye, K. I. Zakariyyah, O. Ajao, O. Olayinka, and H. Sharma, "A Systematic Review of Generative AI for Teaching and Learning Practice," *Education Sciences*, vol. 14, no. 6, p. 636, 2024.
- [11] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proc. 5th Future Business Technology Conf. (FUBUTEC)*, Porto, Portugal, 2008, pp. 5–12.
- [12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.