Auti Kranti Somnath. 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Student Performance Prediction Using Machine Learning: A Review Of Educational Data Mining Methods For Early Intervention

Auti Kranti Somnath[1]; Prof. S. K. Said[2]; Prof. M. C. Kolse[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Predicting student performance is a central application of educational data mining and learning analytics, aimed at forecasting academic outcomes and identifying at-risk students early enough for teachers to intervene. This paper reviews machine learning approaches to the problem. It examines the data sources used, including prior grades, attendance, assignment submissions, learning-management-system activity and demographic factors, and the models applied, from decision trees, random forests and support vector machines to knowledge tracing and deep sequence models. Public datasets such as the Portuguese secondary-school data and the Open University Learning Analytics Dataset are discussed, together with class imbalance, feature engineering, evaluation metrics and fairness concerns. Real-world early-warning systems that combine prediction with student feedback are reviewed. A five-layer architecture comprising data integration, preprocessing, predictive modelling, explanation and intervention is proposed, with research gaps, objectives and a comparison of algorithms. The review concludes that predictive models are most useful when their outputs are explained, checked for bias and paired with supportive interventions that improve retention and personalized learning.

Index Terms—Educational data mining, learning analytics, student performance prediction, early warning system, random forest, knowledge tracing, fairness

I. INTRODUCTION

In many colleges, a student's difficulties become visible only at the end-semester examination, when it is too late to help. Teachers handling large classes cannot track every student's attendance, submissions and test scores in detail. Yet these records, together with activity logs from online learning platforms, often contain early signals of trouble.

Educational data mining applies machine learning to such records. Romero and Ventura [1] reviewed the field and listed performance prediction as one of its most common tasks, while Siemens and Baker [2] described how learning analytics and educational data mining complement each other in turning data into action.

This paper reviews machine learning approaches to student performance prediction, proposes an architecture for an early warning system, and discusses ethical and practical challenges.

II. RELATED WORK

Cortez and Silva [3] predicted Portuguese secondary-school grades using decision trees, random forests [4], neural networks and support vector machines [5], and found that earlier period grades were the strongest predictors, while other variables helped when earlier grades were unavailable. Their dataset is still widely used.

Kuzilek *et al.* [6] released the Open University Learning Analytics Dataset, which links demographics, assessment results and daily clicks in the virtual learning environment for tens of thousands of students, enabling research on early prediction from behaviour.

Hellas *et al.* [7] systematically reviewed hundreds of studies on predicting academic performance and observed large variation in features, definitions of success and evaluation, which makes results hard to compare.

Knowledge tracing models skill mastery over time. Corbett and Anderson [8] introduced Bayesian knowledge tracing for intelligent tutors, and Piech *et*

© 2026 Auti Kranti Somnath, This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

al. [9] proposed deep knowledge tracing with recurrent networks, which predicts whether a student will answer the next question correctly.

Arnold and Pistilli [10] described Course Signals at Purdue, which displayed traffic-light risk indicators to students and instructors and was associated with improved course outcomes, showing the value of pairing predictions with feedback. Baker and Hawn [11] reviewed algorithmic bias in education and warned that models can perform worse for under-represented groups. Table I summarizes these works.

Ref.	Authors	Year	Technique Used	Merits	Demerits
[8]	Corbett and Anderson	1994	Bayesian knowledge tracing	Models skill mastery over time	Simple skill assumptions
[3]	Cortez and Silva	2008	DT, RF, NN, SVM on school data	Prior grades most predictive	Small, single-region dataset
[1]	Romero and Ventura	2010	Review of educational data mining	Mapped tasks and methods	Pre-deep-learning era
[10]	Arnold and Pistilli	2012	Course Signals early warning	Feedback linked to better outcomes	Causality debated
[9]	Piech et al.	2015	Deep knowledge tracing (RNN)	Improved next-answer prediction	Hard to interpret
[6]	Kuzilek et al.	2017	OULAD learning analytics dataset	Rich behavioural data for early prediction	Distance-learning context
[7]	Hellas et al.	2018	Systematic review of prediction studies	Exposed inconsistent evaluation	No new model
[11]	Baker and Hawn	2022	Review of algorithmic bias in education	Documented unequal model performance	Mostly US studies

TABLE I Summary of Literature Survey

III. PRELIMINARIES

1. Prediction Targets

Prediction may target a final grade (regression), pass or fail (binary classification), grade bands (multi-class classification), dropout risk, or mastery of individual skills. Fig. 1 organizes the data sources and model families.

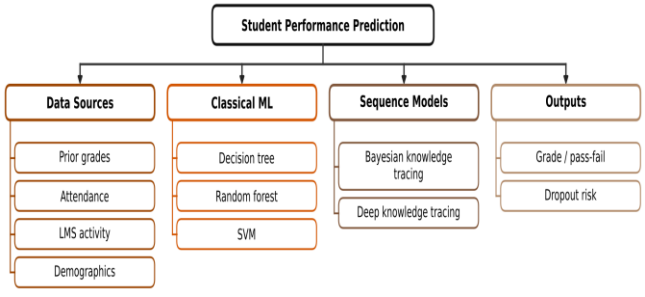


Fig. 1. Data sources and model families in student performance prediction.

2. Types Of Features

Features include demographic and background data, prior academic performance, attendance, assignment timeliness, quiz scores, and behavioural signals such as logins and video views. Early in a course, behavioural features are often the only timely signals available.

IV. PROBLEM STATEMENT AND OBJECTIVES

1. Identified Research Gaps

- **G1:** Many studies predict at the end of a course, too late for intervention.
- **G2:** Results are rarely validated across institutions, so models may not transfer to Indian colleges.
- **G3:** Class imbalance means failing students, the most important group, are often predicted poorly.
- **G4:** Predictions are seldom explained to teachers or students.

- **G5:** Few studies evaluate whether interventions triggered by predictions actually improve outcomes.

2. Objectives Of The Review

- **O1:** To review data sources and machine learning models for student performance prediction.
- **O2:** To compare algorithms in terms of accuracy, interpretability and data needs.
- **O3:** To propose an early warning architecture with explanation and intervention layers.
- **O4:** To examine evaluation, class imbalance and fairness issues.
- **O5:** To identify future research directions.

3. Proposed Framework

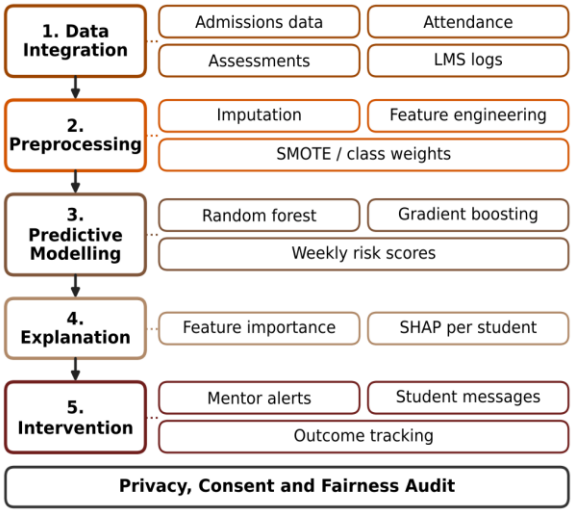


Fig. 2 shows a five-layer early warning system for a college.

Fig. 2. Proposed five-layer early warning architecture.

4. Data Integration Layer

Data from the admissions database, attendance system, internal assessments and the learning management system are combined using student identifiers and stored securely.

5. Preprocessing Layer

Missing values are imputed, categorical variables are encoded, features such as attendance percentage and average submission delay are computed, and class imbalance is handled with SMOTE [12] or class weights.

6. Predictive Modelling Layer

Random forest and gradient-boosting models predict risk categories at several points in the semester, using only data available at each point.

7. Explanation Layer

Feature importance and SHAP values [13] show which factors drive each student's risk, such as low attendance or missed assignments.

8. Intervention Layer

Mentors receive prioritized lists with suggested actions, students receive supportive messages, and outcomes are recorded to evaluate interventions.

V. WORKING METHODOLOGY

Fig. 3 outlines the workflow. Because at-risk students are the minority, models are evaluated with recall and F1-score for the at-risk class rather than accuracy alone:

$$F1 = 2PR / (P + R) \quad (1)$$

where P and R are precision and recall for at-risk students. For fairness, recall is compared across groups such as gender and rural or urban background, and large differences are investigated.

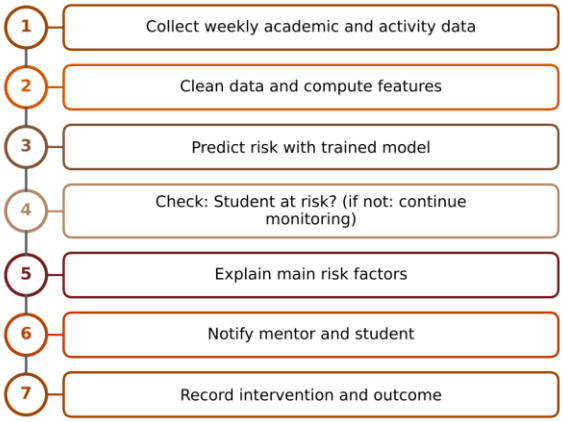


Fig. 3. Workflow of the early warning system.

Predictions are produced weekly so that mentors can act early, and models are retrained each semester.

VI. DISCUSSION

1. Comparison Of Approaches

Table II compares common algorithms. Decision trees are transparent, random forests and gradient boosting

are accurate on tabular data, SVMs perform well on small datasets, and deep knowledge tracing suits fine-grained question-level data.

Algorithm	Accuracy	Interpretability	Data Need
Decision tree	Medium	High	Low
Logistic regression	Medium	High	Low
Random forest	High	Medium	Medium
SVM	High (small data)	Low	Low–Medium
Deep knowledge tracing	High (question level)	Low	High

TABLE II Comparison of Prediction Algorithms

2. Open Challenges And Future Work

Table III summarizes challenges. Future work includes multi-institution datasets for Indian higher education, privacy-preserving analytics, causal evaluation of interventions, and student-facing dashboards that encourage self-regulated learning.

Challenge	Mitigation
Late predictions	Weekly models using early features
Class imbalance	SMOTE, class weights, recall focus
Bias	Group-wise evaluation; remove proxy features
Privacy	Consent, access control, anonymization
Labelling students	Supportive messaging; human mentors decide

TABLE III Challenges and Mitigation

VII. CONCLUSION AND FUTURE WORK

Machine learning can identify students who need help early, when support is still effective. The reviewed literature shows that prior grades and behavioural signals predict outcomes well, that tree ensembles perform strongly on typical college data, and that early warning systems are most valuable when predictions are explained, checked for bias and connected to supportive action.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. 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. Learn. Analytics Knowl. (LAK), 2012, pp. 252–254.
3. P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Future Business Technol. Conf. (FUBUTEC), 2008, pp. 5–12.
4. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
5. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
6. J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open University learning analytics dataset," Sci. Data, vol. 4, Art. no. 170171, 2017.
7. A. Hellas et al., "Predicting academic performance: A systematic literature review," in Proc. ITiCSE Companion, 2018, pp. 175–199.
8. A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," User Model. User-Adapt. Interact., vol. 4, no. 4, pp. 253–278, 1994.
9. C. Piech et al., "Deep knowledge tracing," in Proc. NeurIPS, vol. 28, 2015.
10. K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue: Using learning analytics to increase

student success," in Proc. 2nd Int. Conf. Learn. Analytics Knowl. (LAK), 2012, pp. 267–270.

11. R. S. Baker and A. Hawn, "Algorithmic bias in education," Int. J. Artif. Intell. Educ., vol. 32, pp. 1052–1092, 2022.
12. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
13. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, 2017.