IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, November 2026
Impact Factor: 8.2

Student Performance Prediction Using Machine Learning

Singaldip Pranali Sanjay and Prof. S. J. Ghorpade
Department of Computer Science
K.R.T. Arts, B.H. Commerce and A.M. Science (K.T.H.M.) College, Nashik, India

Abstract: *Predicting academic performance before final examinations can help educators direct support towards students who need it most. This paper presents a research framework for student performance prediction using machine learning, developed from a postgraduate research proposal. The framework combines academic indicators, including previous examination marks, internal assessments, attendance, assignments and study habits, to classify academic risk. It describes data preparation, a comparison of interpretable and ensemble classification methods, and an evaluation plan that emphasizes recall for at-risk students rather than accuracy alone. A teacher-facing reporting interface is proposed to convert predictions into actionable academic support. As model training and institutional validation have not yet been completed, this paper reports a proposed methodology and anticipated contributions, not measured performance.*

Keywords: Student Performance Prediction, Machine Learning, Educational Data Mining, Random Forest, Academic Risk, Learning Analytics

I. INTRODUCTION

Educational institutions routinely record attendance, assessments, assignments and examination outcomes. These records can support more than retrospective grading: they can also help identify patterns associated with future academic difficulties. Traditional monitoring depends heavily on manual interpretation and may identify struggling learners only after a major examination. Educational data mining (EDM) applies analytical methods to such records to support earlier and more consistent intervention [1], [2].

This research investigates how student-level academic and learning-behavior indicators can be organized into a predictive workflow. The intended output is an interpretable risk classification that teachers can review alongside the underlying evidence. The objective is not to replace academic judgment, but to make it easier to prioritize mentoring and remedial support.

II. RELATED WORK AND RESEARCH GAP

Cortez and Silva studied prediction of secondary-school performance from demographic, social and school-related variables, establishing a widely used public dataset for educational prediction research [1]. Reviews of educational data mining describe the application of classification, regression and pattern discovery to learning environments [2]. More broadly, learning analytics research highlights that predictive indicators are most useful when coupled with practical intervention and careful interpretation [3].

A key methodological challenge is the timing of prediction. A model that uses marks recorded close to the final examination may appear strong but be less useful for early intervention. Another challenge is unequal error costs: failing to identify a student who needs support may be more harmful than incorrectly flagging a student for additional review. Therefore, the proposed study emphasizes prediction using information available at the chosen intervention point, transparent evaluation and teacher review.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-41035

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, November 2026
Impact Factor: 8.2

III. PROBLEM STATEMENT AND OBJECTIVES

Problem statement: Student records are often fragmented across attendance registers, internal assessments and previous results. Without a systematic process for combining and analyzing these records, early academic-risk detection remains inconsistent.

The research objectives are to (1) construct a suitable academic dataset; (2) identify informative predictors; (3) compare Decision Tree, Random Forest, K-Nearest Neighbors and Naïve Bayes classifiers; (4) assess risk identification using class-sensitive metrics; and (5) propose a simple reporting interface for educators.

IV. PROPOSED RESEARCH METHODOLOGY

A. Data sources and target definition. The proposed inputs include attendance or absences, first and second internal marks, previous semester results, assignment scores and study-time indicators. A public dataset such as the UCI Student Performance dataset may be used for initial benchmarking [4]. Local institutional data should only be included with appropriate permission and de-identification. The target may be defined as At-Risk versus Not At-Risk using an institution-approved threshold, or as ordered performance bands. Thresholds must be documented before evaluation.

B. Preprocessing. Data quality checks will identify duplicates, invalid marks and missing observations. Numerical features may be imputed and scaled where required; categorical variables may be encoded using training-set mappings. All fitted preprocessing steps must be learned from training folds only, to avoid leakage into validation data. Personally identifying fields should be removed from modeling data.

C. Training and model selection. A stratified train/test split will reserve an untouched test set. Within the training portion, cross-validation will compare a majority-class baseline, Decision Tree, Random Forest, KNN and Naïve Bayes. Hyperparameters will be selected on validation folds rather than the test set. If records span multiple terms for the same students, grouping or chronological splitting should prevent overlap between training and testing.

D. Evaluation. The study will report confusion matrices, precision, recall and F1-score for the At-Risk class, alongside overall accuracy and class distribution. For probability-producing models, calibration and threshold selection may be examined. Feature importance or permutation-based analysis can help explain influential indicators, but association must not be interpreted as causation.

V. PROPOSED SYSTEM ARCHITECTURE

The proposed workflow is: academic records → data validation and preprocessing → trained prediction model → risk classification and explanation → educator dashboard → optional academic intervention. The interface may summarize risk categories, relevant assessment trends and cases requiring human review. Access controls and data minimization should protect sensitive student records.

VI. EXPECTED OUTCOMES AND DISCUSSION

The intended research output is a reproducible comparison of candidate algorithms and a documented process for identifying students who may benefit from support. Previous academic marks are plausible predictors, while attendance and assignment patterns may provide earlier signals; their actual contributions must be established experimentally. A useful model should maintain acceptable recall without producing an impractical number of false alerts. Any dashboard should communicate uncertainty and avoid presenting risk labels as permanent judgments.

No machine-learning model has yet been trained or tested for this study. Accordingly, no dataset-specific accuracy, confusion matrix or comparative ranking is claimed here. These results must be generated through implementation before the work can be presented as a completed empirical study.

VII. LIMITATIONS AND FUTURE SCOPE

Potential limitations include small or unrepresentative samples, missing attendance information, differences in grading practices, data leakage and model bias. Future work includes implementation in Python, comparison on a public

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, November 2026 Impact Factor: 8.2

benchmark and approved college records, fairness checks across relevant groups, and prospective validation of whether early alerts actually improve student outcomes. Integration with a learning management system may later support periodic updates, subject to consent and institutional governance.

VIII. CONCLUSION

This paper outlines a structured and ethically conscious approach to predicting student academic performance with machine learning. It integrates academic features, comparative model assessment, class-sensitive evaluation and educator-facing reporting. The proposed framework provides a foundation for the experimental stage of M.Sc. Computer Science research. Its practical value will depend on verified results, transparent reporting and responsible use in academic decision-making.

ACKNOWLEDGMENT

The authors acknowledge the academic guidance of Prof. S. J. Ghorpade and the Department of Computer Science, K.T.H.M. College, Nashik.

REFERENCES

[1] P. Cortez and A. Silva, “Using Data Mining to Predict Secondary School Student Performance,” in Proc. 5th Future Business Technology Conf., 2008.
[2] 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.
[3] R. S. J. d. Baker and K. Yacef, “The state of educational data mining in 2009: A review and future visions,” J. Educ. Data Mining, vol. 1, no. 1, pp. 3–17, 2009.
[4] P. Cortez, “Student Performance,” UCI Machine Learning Repository, 2014. https://doi.org/10.24432/C5TG7T
[5] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32, 2001.