Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Machine Learning-Based Identification of Students at Risk of Poor Academic Performance: A Comparative Study

Athul Varghese

MCA Graduate; Teacher, De Paul School, Vambori, Maharashtra

Abstract

Identifying students who are at risk of poor academic performance can help educational institutions provide timely academic support and intervention. This study investigates whether machine learning can identify students at risk using demographic, social, behavioral, and educational characteristics. The study uses the UCI Student Performance mathematics dataset containing 395 students. The final mathematics grade (G3) was converted into a binary target: students with G3 below 10 were classified as at risk, while students with G3 of 10 or above were classified as not at risk. First- and second-period grades (G1 and G2) were excluded to avoid direct reliance on previous examination performance and to examine a more useful early-warning setting. Five classification algorithms—Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and K-Nearest Neighbors—were evaluated using stratified five-fold cross-validation. Accuracy, precision, recall, F1-score, and ROC-AUC were used as evaluation measures, while confusion-matrix analysis was used to examine classification errors. Random Forest achieved the highest accuracy (69.11%) and ROC-AUC (69.00%), whereas Logistic Regression produced the highest recall (56.15%) and F1-score (51.23%) among the tested models. Feature analysis identified school support, sex, previous failures, social activity, and absenteeism among the more influential variables.

The findings indicate that machine learning can identify meaningful patterns associated with academic risk, although predictive performance remains moderate and false predictions are substantial. The results also demonstrate that accuracy alone can be misleading when the practical objective is to identify at-risk students. Machine learning may therefore be useful as a supplementary early-warning tool, provided that predictions are interpreted alongside educator judgment, student context, and appropriate academic-support procedures rather than used for automatic decisions.

Keywords: Machine Learning, Student Performance, Academic Risk, Educational Data Mining, Predictive Analytics

1. Introduction

Academic performance is an important indicator of students' educational progress. Identifying students who may experience poor academic outcomes can allow educational institutions to provide additional support before academic difficulties become severe. Traditional identification often relies on examination results, attendance, teacher observations, and individual interactions. Machine learning

offers another approach by learning patterns from multiple student characteristics and producing predictive estimates.

Educational data mining has developed into an important research area for analyzing educational information and predicting student outcomes. Reviews of student-performance prediction research describe a broad range of machine learning and data-mining methods and emphasize the importance of problem formulation, data preprocessing, model validation, and practical interpretation (Zhang et al., 2021).

The UCI Student Performance dataset was collected from two Portuguese schools and contains demographic, social, school-related, and academic attributes. The dataset documentation specifically notes that G1 and G2 are strongly correlated with G3, the final grade, and that prediction without these variables is more difficult but potentially more useful (Cortez, 2008). This observation motivates the present study.

The study asks: Can machine learning identify students at risk of poor academic performance using demographic, social, behavioral, and educational characteristics? The objective is to compare five classification algorithms and determine which model provides the most useful balance for an educational early-warning setting.

2. Related Work

Cortez and Silva (2008) applied data-mining techniques to student-performance data from Portuguese secondary schools and examined classification and regression approaches. Their work established the relevance of demographic, social, school, and academic variables for predicting student outcomes. The UCI repository provides the associated Student Performance dataset and identifies classification and regression as supported tasks (Cortez, 2008).

Zhang et al. (2021) reviewed student-performance prediction from an educational data-mining perspective and emphasized a pipeline involving data collection, preprocessing, problem formulation, model learning, prediction, and result analysis. Their review also highlights limitations such as dataset availability, validation practices, and the gap between predictive models and educational application.

Yağcı (2022) compared several machine learning algorithms for predicting student academic performance and reported that machine learning can support early identification of students at risk of failure. Such studies demonstrate the potential of predictive analytics, while also showing that model performance depends strongly on the available features and prediction setting.

The present study focuses specifically on an early-warning formulation in which G1 and G2 are excluded. Rather than reproducing information already contained in previous grades, the study examines whether other student characteristics can provide useful signals of academic risk.

2.1 Research Objectives

To develop machine learning models for identifying students at risk of poor academic performance.
To compare the predictive performance of Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and K-Nearest Neighbors.

To evaluate the models using accuracy, precision, recall, F1-score, and ROC-AUC. To identify student characteristics that contribute to academic-risk prediction.

2.2 Research Hypotheses

H0: Machine learning algorithms cannot identify students at risk of poor academic performance with meaningful predictive performance using demographic, social, behavioral, and educational characteristics.

H1: Machine learning algorithms can identify students at risk of poor academic performance with meaningful predictive performance using demographic, social, behavioral, and educational characteristics.

3. Methodology

3.1 Dataset

The study uses the mathematics portion of the UCI Student Performance dataset. The uploaded file contains 395 student records and 33 variables. The variables describe demographic, family, school, social, behavioral, and academic characteristics. The final grade G3 ranges from 0 to 20.
The dataset contains no missing values and no duplicate records in the analyzed file.

3.2 Target Definition and Feature Selection

A binary target was constructed from G3. Students with G3 < 10 were classified as At Risk,
while students with G3 ≥ 10 were classified as Not At Risk. The resulting data contained 130 at-risk students (32.91%) and 265 not-at-risk students (67.09%). G1 and G2 were excluded because they represent earlier-period grades and are strongly related to G3. Excluding them makes the task more challenging but better represents an early-warning scenario.

3.3 Preprocessing

Numerical variables were standardized for scale-sensitive algorithms, while categorical variables were one-hot encoded. Preprocessing was implemented within machine-learning pipelines so that transformations were learned from the training portion of each cross-validation fold. This reduces the risk of information leakage between training and validation data.

3.4 Machine Learning Models

Five supervised classifiers were compared: Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). The models were selected to represent linear, kernel-based, ensemble, rule-based, and distance-based approaches.

3.5 Evaluation

Stratified five-fold cross-validation was used because the dataset is relatively small and contains two target classes. Accuracy, precision, recall, F1-score, and ROC-AUC were calculated. Recall is particularly important because an early-warning system should minimize the number of genuinely at-risk students that it fails to identify. Confusion matrices were also examined.

3.6 Feature Analysis

Permutation-based feature importance was used with the Random Forest model to identify variables that contributed to predictive performance. Feature importance was interpreted as predictive relevance rather than causal influence.

4. Results

4.1 Model Comparison

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	64.81%	47.10%	56.15%	51.23%	68.05%
SVM	66.58%	49.21%	47.69%	48.44%	66.74%
Random Forest	69.11%	60.00%	18.46%	28.24%	69.00%
Decision Tree	61.01%	42.59%	53.08%	47.26%	61.16%
KNN	69.37%	58.18%	24.62%	34.59%	64.19%

Table 1. Five-fold cross-validated performance of the classification models.

The results show that no model was best on every metric. KNN produced the highest accuracy (69.37%), closely followed by Random Forest (69.11%). Random Forest also achieved the highest ROC-AUC (69.00%). However, these models had low recall for the at-risk class, particularly Random Forest at 18.46%. Logistic Regression achieved the highest recall (56.15%) and the highest F1-score (51.23%), indicating a better balance between identifying at-risk students and limiting incorrect risk classifications.

Because the primary purpose of the study is early identification of students who may require support, recall and F1-score are more informative than accuracy alone. Therefore, Logistic Regression is selected as the primary model for the early-warning interpretation, while Random Forest is recognized as the strongest model by overall accuracy/ROC-AUC.

4.2 Confusion Matrix

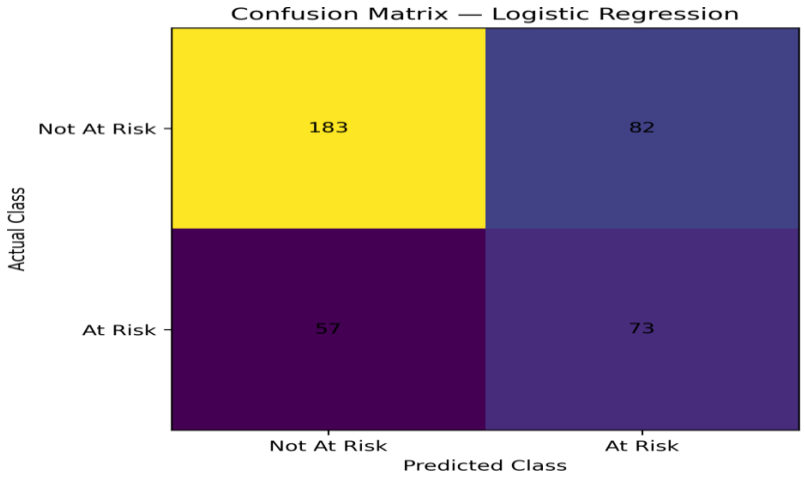

Figure 1. Out-of-fold confusion matrix for the primary Logistic Regression model.

The Logistic Regression model correctly classified 183 of the 265 not-at-risk students and 73 of the 130 at-risk students. It therefore identified 56.15% of the at-risk group, while 57 at-risk students were incorrectly classified as not at risk. The false-negative count demonstrates that the model is not sufficiently reliable to be used as an automatic intervention mechanism.

4.3 Feature Importance

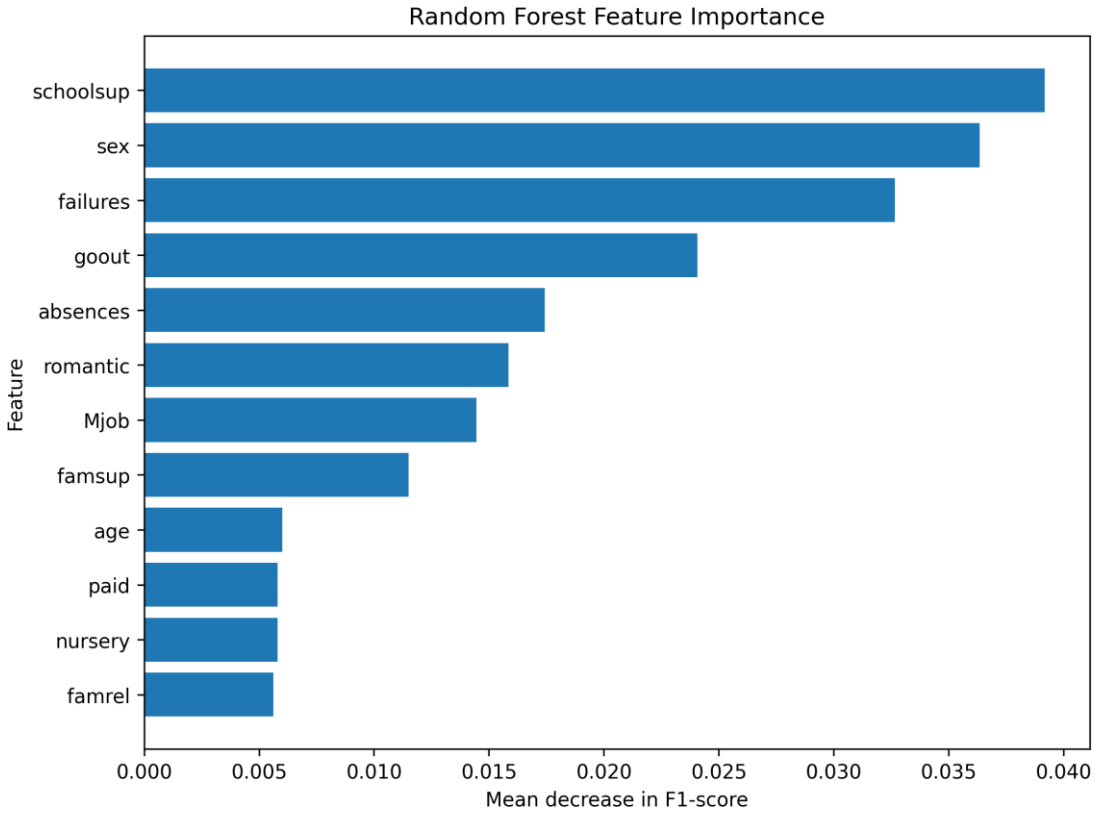

Figure 2. Permutation-based feature importance from the Random Forest model.

The leading variables in the Random Forest permutation analysis included school support, sex, previous failures, going out with friends, and absences. Family and school-related characteristics also appeared among the more influential variables. These results suggest that academic risk is associated with a combination of prior academic history, attendance, social circumstances, and educational support. The analysis does not establish that any of these variables causes poor performance.

5. Discussion

The findings provide moderate evidence that machine learning can identify patterns associated with academic risk without using previous examination grades. This is an important distinction because G1 and G2 are highly informative about G3, and their exclusion creates a more realistic early-warning setting (Cortez, 2008).

Random Forest and KNN achieved higher accuracy than Logistic Regression, but their recall for the at-risk group was considerably lower. In an educational intervention context, this difference matters: a model with high accuracy may still be unsuitable if it systematically fails to identify many students who

need support. Logistic Regression provided the best F1-score and recall among the tested models, making it the most appropriate primary model for the study's stated objective.

The feature analysis also supports the interpretation that academic risk is multifactorial.
Previous failures and absences are intuitively relevant indicators, while school and family support may reflect the wider learning environment. These variables should be treated as signals for further investigation rather than as deterministic labels.

The findings are consistent with the broader educational data-mining literature, which emphasizes that predictive models can support educational decision-making but require careful
validation and interpretation (Zhang et al., 2021; Yağcı, 2022). The present results also demonstrate why evaluation should include class-specific measures rather than accuracy alone.

6. Limitations

The dataset contains only 395 students from two Portuguese schools, limiting generalizability to other educational systems and populations. The binary definition of academic risk is also dependent on the chosen G3 threshold. The study is observational, so feature importance should not be interpreted causally. In addition, excluding G1 and G2 intentionally reduces predictive information. Finally, the models produce classification errors and should not be used to make automatic decisions about individual students.

7. Future Scope

Future work should validate the approach on larger and more diverse datasets collected from multiple institutions. Additional ensemble and boosting methods could be compared, and explainable-AI techniques such as SHAP could be used to provide student-level explanations.

Longitudinal data could also support dynamic risk prediction throughout an academic year. Most importantly, future studies should evaluate whether interventions triggered by model predictions actually improve student outcomes.

8. Conclusion

This study examined whether machine learning can identify students at risk of poor academic performance using demographic, social, behavioral, and educational characteristics. Five classification algorithms were compared using the UCI Student Performance mathematics dataset while excluding G1 and G2. Random Forest achieved the highest overall accuracy and ROC-AUC, whereas Logistic Regression achieved the strongest recall and F1-score for the at-risk class. The findings indicate that machine learning can provide useful signals for academic-risk identification, but predictive performance remains moderate and false predictions are substantial. Therefore, the most appropriate application is as a supplementary early-warning tool that helps educators prioritize further assessment and support, rather than as an autonomous decision system.

References

1. Cortez, P. (2008). Student performance [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

2. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), Proceedings of the 5th Annual Future Business Technology Conference (pp. 5–12). EUROSIS.
3. Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9, Article 11. https://doi.org/10.1186/s40561-022-00192-z
4. Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. Frontiers in Psychology, 12, 698490. https://doi.org/10.3389/fpsyg.2021.698490