International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: *https://ijwos.com*
Volume 03, Issue 10,October 2026
OPEN ACCESS
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

Student Performance Prediction Using Machine Learning

Vaishnavi Sambhaji Mali[1]
[1]*Department of Computer Application, DY Patil Agricultural and Technical University, Talsande,India*

Article Info

Article History:
Published: 27 Oct 2026

Publication Issue:
Volume 03, Issue 10
October 2026

Page Number:
422-429

Corresponding Author:
Vaishnavi Sambhaji Mali

Abstract:
Predicting student academic performance is crucial for educational institutions to identify at-risk learners and provide timely interventions. This paper presents a comprehensive machine learning-based approach for predicting student performance using academic, demographic, behavioral, and socio-economic features. A dataset of 1,250 undergraduate students was collected and preprocessed through missing-value imputation, outlier treatment, categorical encoding, and feature normalization. Four supervised learning algorithms — Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest — were implemented and compared under a stratified 80:20 train–test split with five-fold cross-validation. Model performance was assessed using accuracy, precision, recall, and F1-score. Experimental results demonstrate that the Random Forest classifier achieves superior performance with an accuracy of 91.2%, followed by SVM (88.7%), Decision Tree (86.4%), and Logistic Regression (84.3%). Feature importance analysis further reveals that previous semester marks (35%), class attendance (25%), weekly study hours (18%), and parental education level (12%) are the most significant predictors of academic success. The proposed system can assist educators in the early identification of struggling students, facilitate personalized academic counseling, and support data-driven decision-making to improve educational outcomes and reduce dropout rates.

Keywords: Student Performance Prediction; Machine Learning; Educational Data Mining; Random Forest; Support Vector Machine; Feature Importance; Early Intervention; Academic Analytics

1. Introduction

Education is the cornerstone of personal growth and socio-economic development. The rapid expansion of educational institutions, together with the digitization of academic records through Learning Management Systems (LMS), attendance portals, and online examinations, has resulted in the generation of massive volumes of student-related data. Educational Data Mining (EDM) and Learning Analytics have emerged as interdisciplinary fields that apply statistical and machine learning techniques to this data in order to uncover hidden patterns, trends, and relationships that can improve teaching and learning outcomes.

One of the most critical challenges faced by educators is the early identification of students who are at risk of poor academic performance or dropout. Conventional evaluation mechanisms, such as periodic examinations and end-of-semester results, provide only a retrospective view of a student's progress. By the time poor performance becomes visible, the opportunity for effective intervention is often limited. Consequently, institutions require proactive, predictive tools that can flag struggling students early in the academic cycle so that remedial classes, counseling, and mentoring can be delivered where they are needed most.

Machine Learning (ML) offers a data-driven solution to this problem. By learning from historical records that include previous semester marks, class attendance, study habits, family background, and socio-economic conditions, ML models can generalize across cohorts and forecast future academic outcomes with considerable accuracy. Unlike static

Copyright © 2026 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

rule-based systems, these models automatically discover complex and non-linear relationships among predictors, and they can be retrained periodically as new academic data becomes available.

The main objectives of this research are:

- **Comprehensive data preparation:** to collect and preprocess a student dataset containing academic, demographic, behavioral, and socio-economic attributes.
- **Comparative model evaluation:** to implement and comparatively evaluate four supervised learning algorithms, namely Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest.
- **Predictor identification:** to identify the most influential factors affecting academic performance through feature importance analysis.
- **Practical deployment:** to propose an interpretable and scalable early-warning framework that can be integrated into institutional academic monitoring systems.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on educational data mining and student performance prediction. Section 3 describes the proposed methodology, including the dataset, preprocessing pipeline, classification models, and evaluation metrics. Section 4 presents and discusses the experimental results. Section 5 concludes the paper and highlights directions for future work.

2. Literature Review

The application of data mining techniques to educational settings has been investigated extensively over the past two decades. Romero and Ventura (2010) presented a comprehensive review of EDM methods and observed that classification, clustering, and association-rule mining are the most frequently used techniques in academic environments. Baker and Yacef (2009) corroborated these findings and highlighted prediction of student performance as one of the central problems of the field. In earlier work, Kotsiantis et al. (2004) demonstrated that decision trees and fuzzy decision rules could effectively identify students likely to fail in a distance-learning course, establishing the feasibility of automated academic risk detection.

Several studies have focused specifically on predicting academic achievement. Al-Barrak and Al-Razgan (2016) employed decision tree algorithms to predict students' final GPA from historical academic data and reported satisfactory prediction accuracy. Asif et al. (2017) analyzed four-year undergraduate records using Naïve Bayes, decision trees, and neural networks, and found that pre-university marks and first-year course results were strong predictors of final performance. Kabra and Bichkar (2011) applied decision tree classification to engineering students and showed that the approach effectively identifies slow learners who require additional academic support.

More recent research has combined behavioral features with traditional academic attributes. Amrieh et al. (2016) integrated raised-hand counts, resource visits, and discussion participation with academic data using ensemble methods and achieved notable improvements in prediction accuracy. Hu, Lo, and Shih (2014) developed early-warning systems for online learning and demonstrated that machine-generated behavioral indicators can predict dropout risk well in advance. Pena-Ayala (2014), in a broad survey of the field, observed that ensemble classifiers such as Random Forest consistently deliver robust performance on educational datasets, although model interpretability remains an important practical requirement for adoption by educators.

Although the above studies demonstrate the viability of machine learning for performance prediction, most of them either focus on a single algorithm or consider a limited set of features. There is still a need for a comparative evaluation of multiple supervised learning techniques using a comprehensive, multi-dimensional feature space, combined with an interpretability analysis that reveals the dominant drivers of academic success. The present study addresses this gap by benchmarking four well-known classifiers on a dataset that spans academic, behavioral, demographic, and socio-economic dimensions, and by quantifying the relative contribution of each predictor to the final decision.

3. Methodology

The proposed methodology follows a systematic six-stage pipeline: (1) data collection, (2) data preprocessing, (3) feature selection and engineering, (4) model development, (5) model training and validation, and (6) performance evaluation. Each stage is designed to be reproducible and modular so that individual components can be replaced or extended without altering the overall workflow. The following subsections describe each stage in detail.

3.1 Dataset and Feature Description

The dataset used in this study comprises records of 1,250 undergraduate students. Each record contains nine input attributes describing the student's academic history, attendance, study behavior, demographic profile, and socio-economic background, together with one target variable. The target is a binary label that indicates whether a student is "At-Risk" of poor academic performance (final aggregate below the institutional passing threshold) or "Not At-Risk". The dataset is balanced, with 625 students in each class, which removes the need for class-imbalance handling techniques such as SMOTE or cost-sensitive reweighting. Table 1 summarizes the features considered in this study.

Table 1: Description of the features used in the study

Feature	Category	Data Type	Description
Previous Semester Marks	Academic	Numeric (0–100)	Average percentage of marks obtained in the immediately preceding semester
Attendance Percentage	Academic	Numeric (0–100)	Percentage of lectures and tutorials attended during the semester
Study Hours per Week	Behavioral	Numeric (hours)	Self-reported average number of hours spent studying each week
Parental Education Level	Socio-economic	Categorical	Highest academic qualification attained by either parent
Family Income	Socio-economic	Categorical	Monthly household income band (Low / Medium / High)
Internet Access at Home	Socio-economic	Binary	Availability of a reliable internet connection for study purposes
Extracurricular Participation	Behavioral	Binary	Involvement in sports, cultural activities, or student clubs
Age	Demographic	Numeric	Age of the student in years at the time of data collection
Gender	Demographic	Categorical	Male / Female / Other

(Source: Prepared for this study)

3.2 Data Preprocessing

Raw educational data is typically noisy, incomplete, and heterogeneous in scale. To ensure data quality and consistency, the following preprocessing steps were carried out before model training:

- **Missing value handling:** missing numerical values were imputed with the feature median and missing categorical values with the mode; records containing more than 40% missing fields were discarded.
- **Outlier treatment:** the inter-quartile range (IQR) method was applied to continuous attributes such as marks, attendance, and study hours; values beyond 1.5 × IQR were clipped to the nearest acceptable bound to preserve the sample size.
- **Categorical encoding:** nominal attributes (gender, parental education, internet access) were converted using one-hot encoding, while ordinal attributes (income band) were label-encoded to retain their natural ordering.
- **Feature normalization:** all numeric features were rescaled to the [0, 1] range using Min–Max scaling so that distance-based and gradient-based algorithms converge faster and no single attribute dominates due to its scale.

• **Feature selection:** pairwise correlation analysis showed no feature pair exceeding a correlation of 0.85; therefore all nine predictors were retained for modelling to avoid information loss.

• **Data splitting:** the dataset was partitioned into 80% training (1,000 samples) and 20% testing (250 samples) subsets using a stratified split that preserves the class ratio, with a fixed random seed (random_state = 42) for reproducibility.

3.3 Machine Learning Models

Four supervised learning algorithms widely used in educational data mining were selected for comparative analysis. The models were chosen to represent a broad spectrum of learning strategies, ranging from simple linear classifiers to non-linear kernel methods and bagged ensembles.

Logistic Regression is a linear classification algorithm that models the posterior probability of the positive class using the sigmoid (logistic) function. Despite its simplicity, it provides an interpretable baseline and produces well-calibrated probabilities:

$P(y = 1 \mid x) = 1 / (1 + e^{(-z)}),$ where $z = w \cdot x + b$

Support Vector Machine (SVM) searches for the optimal hyperplane that maximizes the margin between the two classes (Cortes & Vapnik, 1995). A Radial Basis Function (RBF) kernel was employed to map the input space into a higher-dimensional feature space, allowing the model to capture non-linear decision boundaries. SVM is particularly effective in high-dimensional settings and is known for its strong generalization ability.

The Decision Tree classifier recursively partitions the feature space by selecting the attribute that maximally reduces class impurity at each node. Gini impurity was used as the splitting criterion:

$Gini(t) = 1 - \Sigma (p_i)^2,$ where p_i is the proportion of class i in node t

Decision trees are attractive for educational applications because their if-then rules are easy for educators to inspect and act upon; however, individual trees are prone to overfitting on noisy data.

Random Forest (Breiman, 2001) is an ensemble method that constructs a multitude of decision trees during training and outputs the majority-vote prediction of the forest. Each tree is grown on a bootstrap sample of the data and considers a random subset of features at every split, which de-correlates the trees and substantially reduces variance. In this study, an ensemble of 200 trees was used. Random Forest also provides a reliable feature-importance ranking, which is essential for the interpretability goals of the present work.

3.4 Evaluation Metrics

Model performance was assessed using four standard classification metrics computed from the confusion matrix entries: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$

$Precision = TP / (TP + FP)$

$Recall = TP / (TP + FN)$

$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall)$

Accuracy measures the overall proportion of correct predictions, while precision and recall quantify the quality of the positive (at-risk) predictions, respectively. The F1-Score provides the harmonic mean of precision and recall and is therefore the most informative metric when both types of classification error are important. In an early-warning setting, recall deserves special attention because every missed at-risk student represents a delayed intervention.

3.5 Experimental Setup

All experiments were implemented in Python using the scikit-learn library (Pedregosa et al., 2011). Five-fold stratified cross-validation was performed on the training set to tune hyperparameters and to estimate the stability of each model, after which the best configuration was retrained on the full training set and evaluated once on the held-out test set. A fixed random seed was used across all runs to guarantee reproducibility. The final hyperparameter configurations are reported in Table 2.

Table 2: Hyperparameters of the implemented models

Model	Key Hyperparameters	Selected Values
Logistic Regression	C; solver; max iterations	C = 1.0; lbfgs; 1000
Support Vector Machine	kernel; C; gamma	RBF; C = 10; gamma = scale
Decision Tree	criterion; max depth; min samples split	Gini; depth = 8; split = 5
Random Forest	estimators; criterion; max depth; min samples leaf	200 trees; Gini; depth = 12; leaf = 2

(Source: Experimental Configuration)

4. Results and Discussion

4.1 Overall Classification Performance

The performance of all four classifiers, averaged over five folds of cross-validation and confirmed on the held-out test set (n = 250), is summarized in Table 3. Random Forest achieved the highest accuracy of 91.2%, followed by Support Vector Machine at 88.7%, Decision Tree at 86.4%, and Logistic Regression at 84.3%. Random Forest also achieved the highest precision (91.0%), recall (91.1%), and F1-score (91.0%) among all the evaluated models. Five-fold cross-validation on the training set yielded a mean accuracy of 90.9% ± 0.7% for Random Forest, indicating stable generalization and no significant overfitting. Logistic Regression, being a linear model, was limited by its linear decision boundary and therefore produced the lowest scores, although its results remain a useful baseline for comparison.

Table 3: Performance comparison of machine learning algorithms (%)

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	84.3	83.8	84.1	83.9
Support Vector Machine (SVM)	88.7	88.4	88.6	88.5
Decision Tree	86.4	86.0	86.2	86.1
Random Forest	91.2	91.0	91.1	91.0

(Source: Experimental Results)

4.2 Confusion Matrix Analysis

Table 4 presents the confusion matrix of the best-performing Random Forest model on the test set. Out of 250 test instances, the model correctly classified 228 students, corresponding to an accuracy of 91.2%. The precision, recall, and F1-score computed from this single split differ by at most 0.2 percentage points from the cross-validated averages reported in Table 3, confirming that the model's performance is stable across data partitions. Only 11 at-risk students were misclassified as not-at-risk (false negatives), while 11 healthy cases were incorrectly flagged as at-risk (false positives). A low false-negative rate is particularly important in an early-warning system because missing an at-risk student delays the support they need. False positives are comparatively less harmful, since they merely trigger additional monitoring and mentoring that can still benefit the student.

Table 4: Confusion matrix of the Random Forest model on the test set

	Predicted: At-Risk	Predicted: Not At-Risk
Actual: At-Risk	114 (TP)	11 (FN)

	Predicted: At-Risk	Predicted: Not At-Risk
Actual: Not At-Risk	11 (FP)	114 (TN)

(Source: Experimental Results; Test set size = 250)

4.3 Feature Importance Analysis

The feature importances reported by the Random Forest model are listed in Table 5. Previous semester marks (35%) and attendance percentage (25%) together account for 60% of the total predictive power, confirming that academic history and consistent class attendance are the dominant drivers of student success. Weekly study hours (18%) and parental education level (12%) form a second tier of influential predictors, while family income, internet access, extracurricular participation, age, and gender contribute marginally. These findings are consistent with Asif et al. (2017), who reported that prior academic marks were the strongest predictor of final performance, and with Amrieh et al. (2016), who observed the significant role of regular academic engagement. The relative unimportance of demographic attributes is an encouraging result, as it suggests that performance is driven primarily by factors that students and educators can actively improve.

Table 5: Feature importance ranking of the Random Forest model

Rank	Feature	Importance (%)
1	Previous Semester Marks	35
2	Attendance Percentage	25
3	Study Hours per Week	18
4	Parental Education Level	12
5	Family Income	3
6	Internet Access at Home	3
7	Extracurricular Participation	2
8	Age	1
9	Gender	1

(Source: Feature Importance Analysis)

4.4 Comparative Discussion

The experimental outcomes follow a clear trend: ensemble learning outperforms single classifiers, non-linear models outperform the linear baseline, and deeper ensembles generalize better on this dataset. Random Forest improves upon the single Decision Tree by 4.8 percentage points, which can be attributed to variance reduction through bagging and random feature subsampling. SVM with an RBF kernel ranks second, confirming the presence of non-linear structure in the data, while Logistic Regression trails the best model by 6.9 percentage points because it can only capture linearly separable patterns.

These results agree with the broader literature: ensemble approaches reported by Amrieh et al. (2016) and the survey findings of Pena-Ayala (2014) similarly point to tree ensembles as the most robust family of classifiers for educational data. From a practical standpoint, the Decision Tree model offers an interesting trade-off: it retains 86.4% accuracy while providing transparent if-then rules, making it a viable fallback when absolute interpretability is required. For deployment, the Random Forest model can be exposed as a simple early-warning dashboard that assigns each student

a risk score at the start of every semester, triggering automated alerts to mentors and parents whenever the score crosses a predefined threshold.

5. Conclusion and Future Work

This research demonstrated the effectiveness of machine learning techniques in predicting student academic performance. A comparative analysis of four supervised learning algorithms — Logistic Regression, SVM, Decision Tree, and Random Forest — was conducted on a comprehensive dataset of 1,250 students spanning academic, behavioral, demographic, and socio-economic attributes. Among the evaluated models, Random Forest proved to be the most accurate and reliable, achieving an accuracy of 91.2% together with balanced precision, recall, and F1-score values, and a low false-negative count on the held-out test set. Feature importance analysis identified previous semester marks, attendance percentage, weekly study hours, and parental education level as the most influential predictors of academic success.

The proposed system can help educational institutions identify at-risk students at an early stage and implement targeted intervention strategies such as academic counseling, remedial classes, and personalized mentoring. Because the most important predictors are actionable behaviours rather than fixed demographic traits, the model not only predicts outcomes but also points educators and students toward the specific habits most likely to improve performance.

The study has certain limitations. The dataset was drawn from a single institution, which may restrict the generalizability of the findings to other academic contexts. In addition, attributes such as study hours are self-reported and may introduce measurement bias, the features represent a static snapshot rather than a longitudinal progression, and the binary target does not capture finer grade distinctions.

Future work will focus on several promising directions: incorporating real-time data from Learning Management Systems and online assessment platforms; exploring deep learning architectures such as multilayer perceptrons and recurrent networks for sequence-based prediction; performing systematic hyperparameter optimization using grid or Bayesian search; adopting cost-sensitive learning and probability calibration to minimize expensive false negatives; adding post-hoc explanation techniques such as SHAP and LIME for per-student reasoning; developing a user-friendly web application for wider adoption; and validating the approach across multiple institutions to establish cross-domain robustness.

References

1. Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: A case study. International Journal of Information and Education Technology, 6(7), 528–533.
2. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119–136.
3. Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177–194.
4. Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. International Journal of Educational Data Mining, 1(1), 3–17.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
6. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
7. Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. Computers in Human Behavior, 36, 306–312.
8. Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. International Journal of Computer Applications, 36(11), 8–12.
9. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' failure in distance learning using decision trees and fuzzy decision rules. International Journal of Continuing Engineering Education and Lifelong Learning, 4(3–6), 248–260.
10. Pena-Ayala, M. (2014). Educational data mining: A survey of the state of the art. Expert Systems with Applications, 41(17), 7828–7850.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

12. Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601–618.