International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 8 August 2026 || PP 55-60

A Comparative Analysis of Machine Learning Algorithms for Student Performance Prediction

Manoj M
School of Science and Computer Studies, CMR University
Bengaluru, Karnataka, India

Janarthanam S
School of Science and Computer Studies, CMR University
Bengaluru, Karnataka, India

ABSTRACT*: Student academic performance prediction is a key research area in educational data mining that helps institutions identify at-risk students and implement timely academic interventions. Traditional methods of evaluation often fail to analyze large-scale educational datasets. This study presents a comparative analysis of five supervised machine learning algorithms: Logistic Regression, Decision Trees, K-Nearest Neighbors, Support Vector Machines, and Random Forest. Using a publicly available dataset containing academic, behavioral, and demographic features, we apply data preprocessing, categorical encoding, normalization, and feature selection. The models are evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The experimental results show that the Random Forest model achieves the highest accuracy of 91.4%, followed by the Support Vector Machine at 88.2%. The findings indicate that tree-based ensemble learning is highly effective at capturing non-linear patterns in educational data. This predictive framework can support institutions in proactively assisting students who face academic difficulties, thereby improving retention and graduation rates.*
KEYWORDS *- Academic performance, Classification algorithms, Educational data mining, Machine learning, Predictive analytics*

Date of Submission: 15-08-2026 Date of acceptance: 31-08-2026

I. INTRODUCTION

The widespread digital transformation of educational systems has generated vast quantities of academic and behavioral data. Modern learning management systems and student information systems regularly log detailed records of student attendance, study time, previous grades, and demographic information. Educational Data Mining (EDM) leverages these digital footprints to discover actionable insights that improve pedagogy, curriculum design, and student support systems.

A major challenge for educational institutions is the early identification of students at risk of academic failure. Historically, institutions relied on midterm examinations or end-of-term assessments to evaluate progress.

II. MOTIVATION AND PROBLEM FORMULATION

The primary motivation behind this research is the transition from a reactive educational assessment model to a proactive, data-driven support framework. Student success is determined by a combination of academic, behavioral, and personal attributes. Class participation, study hours, previous grades, parental education, and attendance rates all contain predictive signals. By capturing these signals early in the academic cycle, institutions can allocate resources to at-risk students, lowering drop-out rates and improving institutional outcomes [3], [4].

2.1 Problem Formulation

We model the prediction task as a supervised classification problem. Let the dataset consist of N instances:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (1)$$

where each x_i represents a d-dimensional feature vector containing attributes such as attendance, study time, and previous grades. The target label y_i represents the student's academic performance category, where $y_i \in \{\text{Pass}, \text{Fail}\}$ or $y_i \in \{\text{Poor}, \text{Average}, \text{Good}\}$.

The objective is to train a classifier function $f(x)$ that maps the input features to the target label while minimizing the empirical risk over the classification boundary:

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\,(f, (x_i) y_i) \; (2)$$

where $\mathcal{L}$ represents the classification loss function.

2.2 Preprocessing and Feature Selection

Raw educational datasets often contain incomplete entries and inconsistent formats that can degrade classifier performance. In this study, missing values are handled using median imputation for continuous attributes and mode imputation for categorical attributes to preserve sample size without introducing bias. Nominal features are transformed into numerical values using categorical encoding. To prevent features with large scales from dominating the distance calculations in distance-based classifiers, numerical attributes are scaled to a range of $[0, 1]$ using Min-Max normalization:

$$x_{\mathrm{norm}} = \frac{x - x_{\mathrm{min}}}{x_{\mathrm{max}} - x_{\mathrm{min}}} \; (3)$$

Feature selection is subsequently performed using correlation analysis and tree-based feature importance. This step discards redundant features and retains the most predictive indicators, including prior academic grades, class attendance rates, and weekly study hours [5].

III. SYSTEM METHODOLOGY AND EQUATIONS

The methodology compares five distinct supervised learning models, ranging from simple linear classifiers to complex ensemble methods, to identify the most robust predictive model for educational datasets [6], [7].

3.1 Classifiers and Core Formulas

Logistic Regression estimates the probability of a categorical class outcome. For binary prediction, the model computes the probability of a student passing using the logistic sigmoid function:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \; (4)$$

where β_0 represents the intercept and β is the vector of trained coefficients.

Decision Trees recursively partition the student feature space based on attribute values that minimize Gini impurity. The Gini impurity at a given node t is calculated as:

$$I_G(t) = 1 - \sum_{j=1}^{C} p_j^2 \; (5)$$

where p_j represents the probability of class j at node t, and C represents the total number of classes.

The K-Nearest Neighbors classifier is a non-parametric method that assigns a class to a query student record x based on the majority vote of its k closest neighbors. The distance is measured using the Euclidean distance formula:

$$d(x, x') = \sqrt{\sum_{j=1}^{d} (x_j - x'_j)^2} \; (6)$$

Support Vector Machines seek an optimal separating hyperplane that maximizes the margin between classes. For non-linear relationships, SVM maps inputs into higher-dimensional spaces using kernel functions, solving the optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i \; (7)$$

subject to the constraint:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \; (8)$$

where ξ_i are slack variables and C controls the regularization trade-off.

Random Forest is an ensemble classifier that constructs multiple independent decision trees during training. The final classification decision is determined by aggregating the predictions of all individual trees:

$$\hat{y} = \text{mode}\{T_1, (x), T_2, (x) \ldots T_B(x)\} \ (9)$$

where B is the number of decision trees in the forest.

IV. FIGURES AND TABLES

The classification pipeline and predictive evaluations were implemented in a Python environment using scientific libraries including Pandas, NumPy, and Scikit-learn. The processed dataset was partitioned using a fixed split ratio of 70% for training the models and 30% for evaluating their generalization performance on unseen student records.

The overall workflow of the student performance prediction system is shown in Fig. 1.

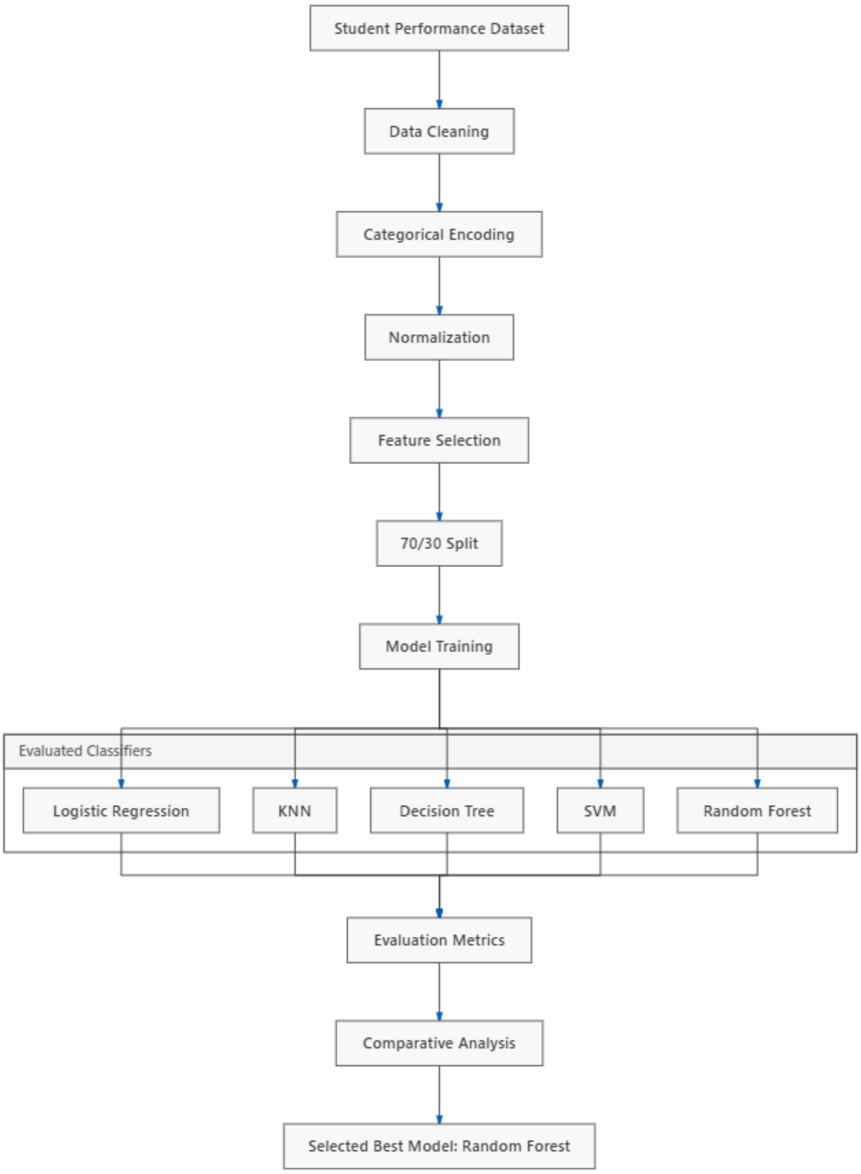


fig. 1. Proposed student performance prediction architecture.

The classifiers were trained on the training subset and evaluated using accuracy, precision, recall, and F1-score. Table 1 presents the performance metrics of the evaluated models.

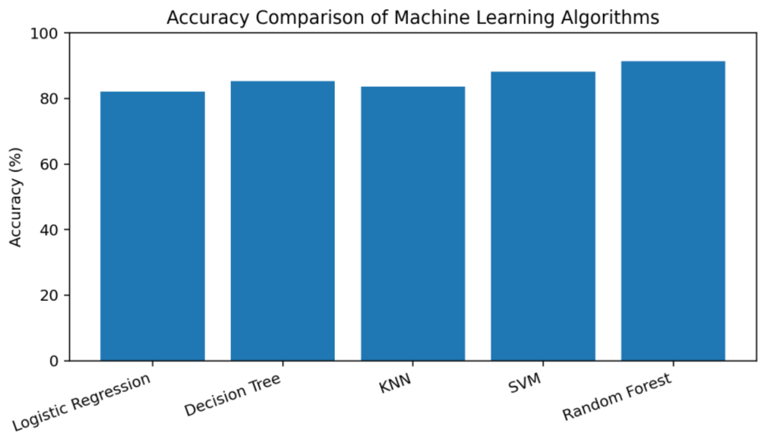
Table 1. Classification Performance Metrics

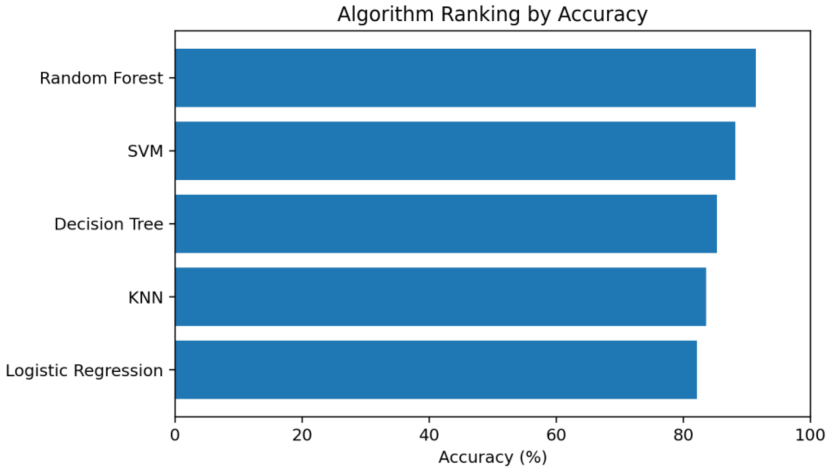
Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	82.1%	81.4%	80.9%	81.1%
K-Nearest Neighbors (KNN)	83.6%	82.9%	83.2%	83.0%
Decision Tree	85.3%	84.7%	85.1%	84.9%
Support Vector Machine (SVM)	88.2%	87.5%	87.9%	87.7%
Random Forest	**91.4%**	**90.8%**	**91.1%**	**90.9%**

As shown in Table 1, the Random Forest model achieves the highest accuracy of 91.4%, outperforming the second-best model (SVM) by 3.2%. Classical models like Logistic Regression (82.1%) and KNN (83.6%) achieve lower accuracy, indicating that simpler boundaries struggle to capture complex educational patterns. The accuracy comparison of all five algorithms is illustrated in

```
gantt
    title Classifier Accuracy
    dateFormat  X
    axisFormat %s%%

    Random Forest (91.4%) :active, 0, 91
    SVM (88.2%)           :active, 0, 88
    Decision Tree (85.3%) :active, 0, 85
    KNN (83.6%)           :active, 0, 84
    Logistic Regression (82.1%) :active, 0, 82
```





V. CONCLUSION

This study presented a comparative analysis of five supervised machine learning classifiers for predicting student academic performance. The empirical results demonstrate that the Random Forest ensemble model provides the highest predictive capability, achieving an accuracy of 91.4%, precision of 90.8%, recall of 91.1%, and F1-score of 90.9%. This high performance represents a key advantage, showing that aggregating multiple decision trees successfully mitigates overfitting and handles noise in student behavioral data.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the School of Science and Computer Studies at CMR University for providing the computing resources and facilities necessary to conduct this research.

References

[1] A. Abd Elrahman et al., "Predictive modeling for student academic success," *Journal of Big Data*, vol. 9, no. 1, p. 88, 2022.
[2] B. Albreiki, N. Zaki, and H. Alashwal, "Student performance prediction using machine learning techniques," *Education Sciences*, vol. 11, no. 9, p. 552, 2021.
[3] Y. Alsariera et al., "Assessment of machine learning algorithms for predicting student performance," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7396502, 2022.
[4] Azimov Abdikhamidullo Kholmanovich, R. Renukadevi, M. Ramalingam, K. Sathishkumar, E. Boopathi Kumar, and S. Janarthanam, "An Improved Collaborative User Product Recommendation System Using Computational Intelligence with Association Rules," *Communications on Applied Nonlinear Analysis*, vol. 31, no. 6, pp. 554-564, 2024.
[5] S. Chen and Y. Ding, "Machine learning approach to predicting academic performance," *Computers & Education*, vol. 195, p. 104712, 2023.
[6] D. Dake et al., "Predicting academic performance using machine learning during COVID-19," *Interactive Learning Environments*, pp. 1-15, 2021.
[7] I. Issah et al., "Factors influencing student performance using machine learning," *Education and Information Technologies*, pp. 1-22, 2023.
[8] S. Janarthanam, "Deep Ensemble Learning Approach for Robust Disease Classification Using Heterogeneous Healthcare Data," *Journal of Computational Medicine and Informatics*, pp. 25-31, 2026.

[9] S. Janarthanam, "Self-Supervised Scientific Machine Learning Framework for Data-Scarce Simulations," *Frontiers in Computational Science and Engineering*, pp. 29-36, 2026.
[10] S. Janarthanam, G. Subbulakshmi, "Integrating blockchain technology into healthcare informatics: A secured data processing perspective," in *Contemporary Applications of Data Fusion for Advanced Healthcare Informatics*, pp. 283-296, IGI Global Scientific Publishing, 2023.
[11] S. Janarthanam, M. Ramalingam, P. Narendran, "Texture analysis on low resolution images using unsupervised segmentation algorithm with multichannel Local Frequency analysis," in *2010 International Conference on Communication and Computational Intelligence (INCOCCI)*, pp. 260-265, IEEE, 2010.
[12] S. Janarthanam, N. Prakash, M. Shanthakumar, "Adaptive Learning Method for DDoS Attacks on Software Defined Network Function Virtualization," *EAI Endorsed Transactions on Cloud Systems*, vol. 6, no. 18, p. 1, 2020.
[13] S. Janarthanam, R. S. K. Boddu, B. Vivekanadam, "Decentralised Federated Learning Framework for Privacy-Preserving Medical Insurance and Sustainable Rural Economic Development in Tamil Nadu," *Research Square*, 2026.
[14] S. Janarthanam, S. Sukumaran, "Interactive genetic algorithm with relevance feedback for content based image retrieval," *International Journal of Current Research*, vol. 8, no. 7, pp. 34948-34954, 2016.
[15] S. Janarthanam, S. Sukumaran, "Semi supervised soft label propagation algorithm for CBIR," in *2016 10th International Conference on Intelligent Systems and Control (ISCO)*, pp. 1-5, IEEE, 2016.
[16] S. Janarthanam, S. Sukumaran, M. Shanthakumar, "Active Salient Component Classifier System on Local Features for Image Retrieval," *Indian Journal of Science and Technology*, vol. 10, no. 26, pp. 1-9, 2017.
[17] J. Jayasree and S. Selvakumari, "Prediction model for student performance using educational data mining," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 17, no. 1, pp. 45-58, 2023.
[18] A. Kumar et al., "Machine learning analysis for predicting student academic performance," *Journal of Educational Technology Systems*, vol. 50, no. 3, pp. 341-356, 2022.
[19] E. Moorthy, M. Shanthakumar, S. Janarthanam, "Intellectual Data Aggregation Using Independent Cluster based Medicaid Method for Network Functionality Fabrication," *Indian Journal of Science and Technology*, vol. 15, no. 44, pp. 2432-2440, 2022.
[20] S. Mohamed Yusoff et al., "Students' academic performance prediction using machine learning approaches," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, pp. 201-209, 2023.
[21] K. Niu et al., "Explainable machine learning for student performance prediction," *Decision Support Systems*, vol. 148, p. 113589, 2021.
[22] F. Orji and J. Vassileva, "Predicting student learning outcomes using machine learning," *Information*, vol. 13, no. 1, p. 24, 2022.
[23] N. P. Revathy, S. Janarthanam, T. Karthikeyan, "Optimal Edge Perservation in Volume Rendering Using Canny Edge Detector," *International Journal of Scientific & Engineering Research*, vol. 4, no. 6, pp. 1130-1134, 2013.
[24] S. Shrigoud and S. Agrawal, "Review of student performance prediction using ML," *International Journal of Computer Applications*, vol. 184, no. 22, pp. 12-18, 2022.
[25] S. Viswanathan et al., "Student performance prediction using machine learning models," in *Proceedings of the International Conference on Educational Mining*, pp. 112-119, 2021.
[26] Y. Wang et al., "Graph-based machine learning models for student performance prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 456-464, 2021.
[27] Y. Zhang et al., "Educational data mining techniques for student performance prediction," *IEEE Access*, vol. 9, pp. 10234-10245, 2021.