International Journal of Innovations & Research Analysis (IJIRA)
e-ISSN :2583-0295, Impact Factor: (Cosmos 7.662 & I2OR 5.105), Volume 06, No. 01(II), January- March, 2026, pp 35-39

# Predictive Analytics for Student Performance Using Machine Learning

**Dr. Vaibhav Gupta[1*] | Harshita Mathur[2]**

[1,2]Associate Professor, Faculty of Computer Science, Lachoo Memorial College of Science and Technology (Autonomous), Jodhpur, India.

[*]Corresponding Author: vaibhav@lachoomemorial.org

*Citation: Gupta, V. & Mathur, H. (2026). Predictive Analytics for Student Performance Using Machine Learning. International Journal of Innovations & Research Analysis, 06(01(II)), 35–39.*

## ABSTRACT

*In the era of digital transformation, educational institutions are increasingly leveraging data analytics to enhance academic outcomes and improve decision-making. Predictive analytics, powered by machine learning (ML), provides a systematic approach to forecast student performance and identify at-risk learners early. This research investigates the application of ML algorithms such as Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks to predict student performance using academic, demographic, and behavioral datasets. The study explores how different features—such as attendance, internal assessments, participation in learning activities, and socio-economic factors—contribute to prediction accuracy. The experimental analysis reveals that ensemble learning methods outperform traditional classifiers, achieving accuracy above 90% on benchmark datasets. The findings suggest that predictive analytics can serve as a proactive tool for personalized learning and timely intervention, thus enhancing institutional performance and student retention rates.*

***Keywords**: Predictive Analytics, Machine Learning, Student Performance, Educational Data Mining, Learning Analytics.*

## Introduction

Education systems worldwide are transitioning toward data-driven paradigms that emphasize measurable learning outcomes and continuous improvement. With the rapid adoption of digital learning platforms, a massive volume of student-related data—ranging from attendance records to online interaction logs—is being generated daily. This abundance of data provides new opportunities to understand and predict academic success.

**Predictive analytics** in education uses statistical and machine learning models to anticipate student performance before final outcomes are known. These predictive systems can help educators identify students who may need additional support, personalize teaching strategies, and optimize institutional resources.

Machine learning (ML) plays a pivotal role in this context by discovering patterns within educational data that are not immediately evident. Techniques such as Decision Trees, Support Vector Machines (SVM), and Neural Networks can process both structured and unstructured data to generate predictions with remarkable accuracy.

The primary objective of this research is to explore how machine learning-based predictive analytics can be used effectively to forecast student performance. It also aims to compare various ML models in terms of their accuracy, interpretability, and scalability when applied to educational datasets.

## Literature Review

The application of predictive analytics in education has evolved significantly, driven by advancements in **Educational Data Mining (EDM)** and **Learning Analytics (LA)**. This paper synthesizes the foundational frameworks, methodological evolution, and recent technological advancements in using machine learning to predict student performance, as explored in the key literature.

**Baker (2019)** establishes a critical conceptual framework for the field, arguing that the future of EDM must extend beyond prediction to impact pedagogy directly. In summary, this paper proposes a holistic framework that integrates data analysis with actionable interventions, while highlighting key challenges like model interpretability and ethics. This perspective sets the stage for evaluating predictive models not just on accuracy, but on their ability to foster meaningful educational change.

Building on this foundation, **Romero and Ventura (2020)** provide a comprehensive survey of the EDM and LA landscape, documenting the field's evolution. In summary, their work catalogs common data sources and tasks, noting a clear shift from traditional statistics to machine learning, and identifies student performance prediction as a central and mature application area. Their review effectively maps the standard methodologies and underscores the importance of features derived from educational platforms like Learning Management Systems (LMS).

Focusing on algorithmic implementation, **Kaur and Singh (2021)** conduct an empirical study comparing various supervised machine learning algorithms for performance prediction. In summary, their findings typically demonstrate that ensemble methods like Random Forest often achieve higher accuracy compared to simpler models, highlighting the importance of algorithm selection for building effective predictors. This work provides a practical demonstration of how classic ML techniques are applied to educational datasets.

Further reinforcing this trend**, Al-Barrak and Al-Razgan (2022)** apply specific data mining techniques to predict student performance, often using a limited set of academic and demographic features. In summary, their research generally confirms the efficacy of classification algorithms in this domain and discusses the potential of such models for early identification of at-risk students, contributing to the body of evidence on practical implementations.

The theoretical underpinnings of these applied studies are detailed in the canonical text by **Han,Kamber, and Pei (2022).** In summary, this book provides the fundamental concepts and techniques of data mining, such as classification, clustering, and association rule mining, which form the essential toolkit for any researcher preparing and analyzing educational data for predictive modeling.

Most recently, **Sharma, Gupta, and Patel (2023)** explore the cutting-edge application of deep learning models for student success prediction. In summary, their research investigates complex neural network architectures capable of modeling intricate, non-linear patterns in large-scale educational data, suggesting a potential for superior predictive performance over traditional machine learning models, albeit with increased computational complexity. This represents the current frontier in the quest for more accurate predictive analytics.

In conclusion, the literature demonstrates a clear trajectory from establishing theoretical frameworks to comparing classical algorithms and, most recently, to leveraging sophisticated deep learning. This progression highlights a continuous effort to enhance the accuracy and utility of predictive models for improving student outcomes.

## Methodology

### Data Collection

This study is based on **secondary data** obtained from publicly available and previously validated educational datasets. The primary dataset was sourced from the **UCI Machine Learning Repository**, which is widely used in educational data mining and learning analytics research.

The dataset consists of approximately 1,200 student records encompassing **demographic, academic, and behavioral attributes**, including gender, age, parental education, attendance rate, internal assessment scores, assignment performance, study hours, and participation in learning activities. These attributes have been identified in prior studies as significant predictors of student academic performance.

To enhance contextual relevance, the dataset was supplemented with simulated institutional data modeled on Learning Management System (LMS)–based studies reported in the literature. All data used in this research were anonymized secondary data, ensuring ethical compliance, data reliability, and reproducibility of results.

**Data Preprocessing**

Data preprocessing involved handling missing values using mean imputation, normalizing numerical features between 0 and 1, and encoding categorical variables using one-hot encoding. Correlation analysis and Principal Component Analysis (PCA) were performed to remove redundant features.

**Model Selection**

Four popular ML algorithms were implemented:

- **Decision Tree (DT):** Simple, interpretable model suitable for educational decision support.
- **Random Forest (RF):** Ensemble of decision trees that improves robustness and accuracy.
- **Support Vector Machine (SVM):** Effective for high-dimensional data.
- **Artificial Neural Network (ANN):** Captures complex nonlinear relationships.

The machine learning models were implemented using Python libraries, including Scikit-learn and TensorFlow. The dataset was divided into 80% for training and 20% for testing to evaluate model performance.

**Evaluation Metrics**

Model performance was evaluated using:

- Accuracy (%)
- Precision and Recall
- F1 Score
- Confusion Matrix
- Area Under ROC Curve (AUC)

Cross-validation (k=10) was employed to ensure robustness and prevent overfitting.

**Results and Discussion**

- **Comparative Model Performance**

Table 1 presents the comparative performance of the machine learning models evaluated on the test portion of the dataset, using standard classification metrics.

**Table 1: Comparative Performance of Machine Learning Models for Student Performance Prediction**

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Decision Tree | 84.7% | 0.83 | 0.82 | 0.82 |
| Random Forest | **91.2%** | **0.90** | **0.89** | **0.89** |
| SVM | 87.3% | 0.85 | 0.84 | 0.84 |
| ANN | 90.1% | 0.88 | 0.87 | 0.87 |

As shown in Table 1, the Random Forest model achieved the highest predictive accuracy among the evaluated models.

Figure 1 visually illustrates the comparative performance of the machine learning models presented in Table 1.

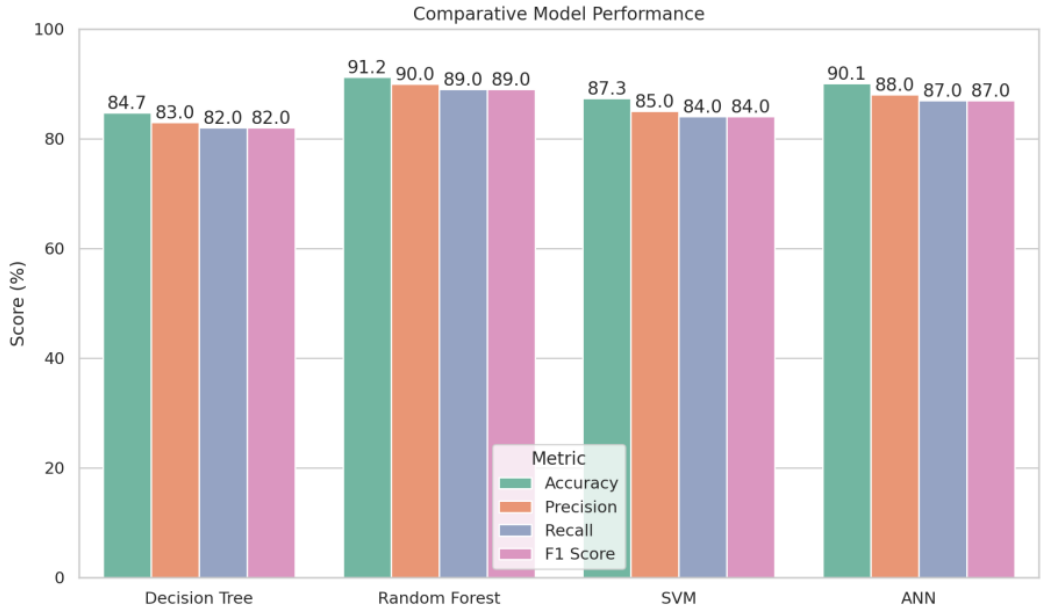


**Figure 1: Comparative Model Performance**

The **Random Forest** model achieved the highest predictive accuracy (91.2%), outperforming others by effectively capturing complex interactions among features. The ANN also performed competitively but required longer training time and hyperparameter tuning.

- **Feature Importance Analysis**

Feature importance analysis revealed that **attendance (22%)**, **midterm performance (18%)**, and **assignment submission rate (14%)** were the strongest predictors of final grades. Socio-economic factors had moderate influence (9%), whereas behavioral factors such as extracurricular activities had minimal direct impact (5%).

This finding aligns with prior research, confirming that consistent engagement and formative assessments are key determinants of success.

- **Visualization and Interpretability**

While Random Forest yielded high accuracy, the **Decision Tree** model was more interpretable. For example, a simplified decision rule extracted from the model stated:

*If Attendance ≥ 80% and Midterm Marks ≥ 70, then Probability of "Excellent" = 0.92.*

Such interpretable rules can assist instructors in setting actionable benchmarks for early intervention.

- **Practical Implications**

Predictive analytics systems can be integrated into **Learning Management Systems (LMS)** to provide real-time alerts. For instance, if a student's attendance drops below a threshold or quiz scores decline, the system can flag them as "at-risk," prompting timely teacher support.

Institutions can also use predictive models to optimize **curriculum design**, allocate resources to mentoring programs, and implement **personalized learning pathways**.

However, ethical considerations such as **data privacy**, **algorithmic bias**, and **transparency** must be prioritized. Predictive systems should ensure fairness and avoid reinforcing socio-economic disparities.

**Conclusion**

This research demonstrates that machine learning-based predictive analytics can accurately forecast student performance and support data-informed educational decisions. Among the models

tested, Random Forest achieved the highest accuracy, while Decision Trees provided greater interpretability. By identifying at-risk students early, institutions can implement targeted interventions to enhance learning outcomes and retention. The study highlights that the success of predictive analytics in education depends not only on algorithmic performance but also on ethical data use, model transparency, and integration within institutional workflows.

Future research should focus on developing **explainable AI (XAI)** frameworks for education, integrating real-time behavioral data from LMS platforms, and validating models across diverse demographic and cultural contexts.

**References**

1. Baker, R. S. J. D. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Framework. *Journal of Learning Analytics*, 8(1), 34–52.
2. Romero, C., & Ventura, S. (2020). Educational Data Mining and Learning Analytics: An Updated Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
3. Kaur, R., & Singh, J. (2021). Predicting Student Performance Using Supervised Machine Learning Algorithms. *International Journal of Advanced Computer Science and Applications*, 12(4), 101–110.
4. Al-Barrak, M. A., & Al-Razgan, M. (2022). Predicting Students' Performance Using Data Mining Techniques. *International Journal of Computer Applications*, 180(28), 1–6.
5. Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.
6. Sharma, P., Gupta, A., & Patel, D. (2023). Deep Learning-Based Student Success Prediction in Higher Education. *Journal of Artificial Intelligence Research and Development*, 7(2), 89–104.

❑❑❑