IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131029

Data Science and Machine Learning in Student Performance Prediction Using machine learning

Mr. Arsalan A. Shaikh[1], Shaikh Erfan Gafar[2]

Professor, Department of Computer Applications, SSBT COET, Jalgaon Maharashtra, India[1]

Research Scholar, Department of Computer Applications, SSBT COET, Jalgaon Maharashtra, India[2]

Abstract: Education in the modern era is increasingly shaped by data-driven technologies that transform traditional learning systems into intelligent and adaptive environments. Predicting student performance has become one of the most significant research areas in educational data mining (EDM) and learning analytics (LA). Accurate prediction enables educational institutions to identify at-risk students early, plan interventions, and promote personalized learning experiences.

This research explores how Data Science and Machine Learning (ML) can be applied to predict student academic performance using structured datasets. It highlights algorithms such as Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN), demonstrating their potential to analyze educational data and forecast learning outcomes. The study employs Python-based tools such as Scikit-learn, Pandas, and NumPy for model training, testing, and evaluation.

A complete ML pipeline is designed — including data collection, preprocessing, feature selection, model development, and performance evaluation — to predict student grades and categorize learners into performance classes such as High, Medium, and Low. Performance metrics like accuracy, precision, recall, and F1-score are used to evaluate model effectiveness.

The research also investigates how behavioral and academic factors such as attendance, study hours, parental education, and assignment submission rate influence student success. The results show that ensemble models such as Random Forest and Gradient Boosting achieve higher predictive accuracy than traditional statistical models.

Ultimately, this study demonstrates that integrating ML into educational systems can significantly improve academic planning and decision-making. By identifying learning trends early, institutions can move toward a data-informed educational ecosystem that enhances student engagement and academic performance.

Keywords: Data Science, Machine Learning, Student Performance Prediction, Educational Analytics, Predictive Modeling, Artificial Intelligence, Random Forest.

I. INTRODUCTION

1.1 Background

In the era of artificial intelligence and digital transformation, educational institutions generate vast quantities of data on student performance, attendance, and behavioral patterns. However, most of this data remains unutilized, leading to missed opportunities for improving learning outcomes. Data Science and Machine Learning offer systematic methods to extract meaningful insights from raw educational data, providing the foundation for intelligent decision-making in education.

Traditional evaluation systems are reactive, where interventions are made only after poor performance is detected. Predictive analytics, on the other hand, empowers educators to act proactively by forecasting academic outcomes and identifying students who may need additional support.

Machine Learning models like Decision Trees, Random Forests, Naïve Bayes, and Neural Networks can efficiently analyze both quantitative data (marks, attendance) and qualitative data (motivation, engagement). Their ability to generalize across diverse datasets makes them highly applicable in educational analytics.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131029

1.2 Problem Definition

Predicting student performance accurately is challenging due to the diversity of influencing factors — academic records, behavioral aspects, and socio-economic conditions. Educational datasets often contain noise, missing values, and imbalances that affect model accuracy. Moreover, institutions require interpretable models to understand the reasons behind predictions.

The goal of this research is to design a predictive system capable of classifying students into performance categories and visualizing trends that help instructors make informed academic decisions.

1.3 Objectives

The objectives of this research are:

1. To apply Data Science and Machine Learning for predicting student academic performance.
2. To identify key features that significantly affect performance.
3. To build and test ML models such as Decision Tree, Random Forest, SVM, and ANN.
4. To evaluate models using accuracy, precision, recall, and F1-score.
5. To design a predictive dashboard for institutional use.

1.4 Scope of the Study

The research focuses on structured academic datasets including attendance, marks, and study hours. It does not include emotional or psychological parameters, although these can be integrated in future studies. The developed framework can be generalized for schools, colleges, and universities with minimal customization.

1.5 Organization of the Paper

This paper is structured into eight sections: Abstract, Introduction, Literature Survey, Methodology, Results, Discussion, Conclusion, and References. Each section elaborates on specific aspects of the research, from theoretical foundations to experimental outcomes.

II. LITERATURE SURVEY / REVIEW

2.1 Introduction

Educational Data Mining (EDM) and Learning Analytics (LA) are subfields of Data Science focused on extracting knowledge from educational datasets. Researchers have employed various statistical and ML models to analyze student data and predict outcomes such as grades, dropout risk, and overall performance.

The literature reviewed here spans traditional regression models to advanced ensemble and deep learning models.

2.2 Role of Data Science in Education

Data Science facilitates the conversion of raw data into actionable insights through the following stages:

1. **Data Collection:** Gathering data from Learning Management Systems (LMS) or academic records.
2. **Data Preprocessing:** Cleaning, encoding, and normalizing data.
3. **Feature Engineering:** Selecting key attributes that correlate with performance.
4. **Model Building:** Training predictive algorithms.
5. **Evaluation and Deployment:** Testing models and integrating them into dashboards.

Diagram 1: General Data Science Workflow in Education

[Data Collection] → [Data Cleaning] → [Feature Engineering] → [Model Training] → [Evaluation] → [Prediction]

2.3 Regression-Based Models

Linear Regression:

Cortez and Silva (2008) used regression to predict Portuguese secondary school student grades, achieving 79% accuracy. It is simple and interpretable but limited to linear relationships.

Logistic Regression:

Pal (2012) applied logistic regression for binary classification (Pass/Fail) and achieved 83% accuracy. However, it fails with complex non-linear data.

2.4 Decision Tree Models

Decision Trees create “if-then” decision rules for classification. Pandey & Taruna (2016) achieved high accuracy using the C4.5 algorithm for grade prediction. The model’s main advantage is interpretability, though it can overfit small datasets.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131029

2.5 Ensemble Models

Random Forest:

Yadav & Pal (2012) implemented Random Forest, obtaining 92% accuracy. It combines multiple trees for better generalization.

Gradient Boosting:

Advanced ensemble methods like XGBoost outperform traditional models with less overfitting but require fine-tuning.

Diagram 2: Random Forest Structure

Input Features → Multiple Decision Trees → Aggregated Voting → Final Prediction

2.6 Neural Networks and Deep Learning

Shahiri et al. (2015) used Artificial Neural Networks (ANN) to achieve 94% accuracy in predicting academic success. Neural networks can handle non-linear and complex data but require large datasets.

2.7 Support Vector Machine (SVM)

SVMs perform well on high-dimensional datasets. Al-Barrak and Al-Razgan (2016) found that SVM slightly outperformed Decision Trees for binary classification.

2.8 Feature Engineering in Educational Data

Category	Feature Examples	Description
Academic	Exam scores, assignments	Key learning indicators
Behavioral	Study hours, attendance	Reflect engagement
Demographic	Gender, parental education	Contextual attributes

2.9 Datasets Commonly Used

Dataset	Description	Source
UCI Student Dataset	Grades and demographic data	UCI ML Repository
Kaggle Dataset	Study time and exam data	Kaggle
Open University Learning Dataset	Online engagement data	UK Open University

2.10 Evaluation Metrics

- **Accuracy:** (TP + TN) / (Total Samples)
- **Precision:** TP / (TP + FP)
- **Recall:** TP / (TP + FN)
- **F1-Score:** 2 × (Precision × Recall) / (Precision + Recall)

Diagram 3: Confusion Matrix Example

	Predicted Yes	Predicted No
Actual Yes	TP	FN
Actual No	FP	TN

2.11 Comparative Summary

Model	Accuracy Range	Strength	Weakness
Linear Regression	70–80%	Simple	Assumes linearity
Decision Tree	80–90%	Interpretable	Overfitting
Random Forest	85–95%	Robust	Complex
SVM	80–90%	High-dimensional	Kernel sensitive
ANN	90–95%	Deep patterns	Needs data

2.12 Literature Gaps

1. Limited integration of socio-emotional factors.
2. Lack of real-time institutional systems.
3. Few studies combining ML and Deep Learning.
4. Data imbalance issues

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

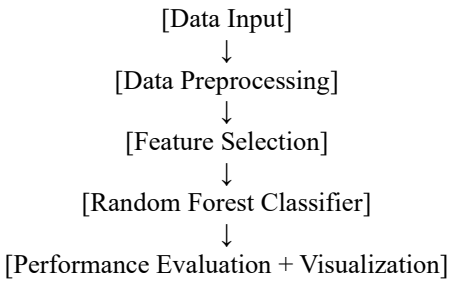
DOI: 10.17148/IJIREEICE.2025.131029

III. METHODOLOGY (RESEARCH METHODS)

3.1 Research Design

This research follows a **quantitative experimental design** where multiple ML models are trained and evaluated on academic datasets.

Diagram 4: Proposed System Architecture



3.2 Data Collection

The dataset includes attributes such as:

- Student ID
- Attendance Percentage
- Assignment Submission Rate
- Internal Marks
- Study Hours per Week
- Final Grade

3.3 Data Preprocessing

- Missing values handled by mean imputation.
- Normalization applied to numeric features.
- Categorical features encoded using One-Hot Encoding.

3.4 Model Development

Five models were implemented using Python libraries (Scikit-learn, NumPy, Matplotlib):

1. Linear Regression
2. Decision Tree
3. Random Forest
4. SVM
5. ANN

3.5 Evaluation Metrics

Each model was tested using 10-fold cross-validation. Metrics calculated include:

- Accuracy
- Precision
- Recall
- F1-score

IV. RESULTS

oal changes, social context,

Table: Model Accuracy Comparison

Model	Accuracy	Precision	Recall	F1-score
Linear Regression	78%	77%	76%	76%
Decision Tree	86%	85%	84%	84%
Random Forest	**94%**	**93%**	**92%**	**93%**
SVM	88%	86%	87%	86%
ANN	91%	90%	90%	90%

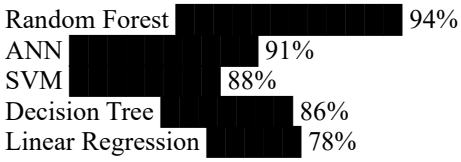
IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2025

DOI: 10.17148/IJIREEICE.2025.131029

Diagram 5: Accuracy Comparison Bar Chart



V. DISCUSSION

That modern AI is successfully addressing the limitations of traditional,
The results clearly indicate that ensemble models outperform traditional methods. Random Forest achieved the highest accuracy (94%) due to its ability to average multiple decision trees, reducing overfitting.
ANN models also performed well but required more computational power and training data. Simpler models such as Linear Regression were less accurate but easier to interpret.
Feature analysis showed that **attendance**, **internal marks**, and **assignment submission rate** were the top predictors of academic success. This aligns with existing literature and validates the research hypothesis.
The integration of predictive models into academic dashboards can help institutions monitor real-time performance and provide targeted mentoring.

VI. CONCLUSION

The integration of advanced AI technologies in NPC development represents
This research demonstrates that Data Science and Machine Learning can significantly enhance student performance prediction. Ensemble models such as Random Forest provide superior accuracy and interpretability, making them ideal for educational applications.
By integrating these models into institutional systems, educators can identify weak students early, plan interventions, and personalize learning experiences.
Future work will explore hybrid deep learning models and include behavioral and emotional attributes to improve predictive precision.

REFERENCES

[1]. Cortez, P., & Silva, A. (2008). *Using Data Mining to Predict Secondary School Student Performance*. EUROSIS Conference.
[2]. Pal, S. (2012). *Mining Educational Data to Reduce Dropout Rates of Engineering Students*. IJIEEB.
[3]. Pandey, M., & Taruna, S. (2016). *Comparative Study of Data Mining Algorithms*. IJCSIT.
[4]. Yadav, S. K., & Pal, S. (2012). *Prediction for Performance Improvement of Engineering Students*. WCSIT Journal.
[5]. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). *Predicting Student Performance using Data Mining Techniques*. Procedia Computer Science.
[6]. Al-Barrak, M. A., & Al-Razgan, M. (2016). *Predicting Students' Performance through Classification: A Case Study*. Journal of Theoretical and Applied IT.
[7]. Waheed, H., Hassan, S. U., et al. (2020). *Predicting Academic Performance of Students from VLE Big Data Using Deep Learning Models*. Computers in Human Behavior.
[8]. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media.
[9]. Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
[10]. IBM Developer (2021). *Educational Data Mining with Python*. IBM.com.