e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

PREDICTIVE MODELS FOR STUDENT PERFORMANCE: A MACHINE LEARNING PERSPECTIVE

Roshan Renji*1, Dr. R. Mahalakshmi*2

*1Research Scholar, Computer Science, Veles Institute of Science and Technology and Advanced Studies, VISTAS, Pallavaram, Chennai, India.

*2Associate Professor, Department of Advanced Computing and Analytics, School of Computing Sciences VISTAS, Chennai, India.

ABSTRACT

Machine learning techniques are increasingly being applied across various domains, with education emerging as a key area of focus. The application of these methods in educational research is growing rapidly, enabling the discovery of underlying patterns in student performance. This research aims to develop a predictive model for academic outcomes using several machine learning classification techniques, such as K-Nearest Neighbor, Decision Tree, Support Vector Machines, Random Forest, and Gradient Boosting. The model considers various factors, including living environment, parent relationships, educational background, employment status, backlogs, attendance, internet availability, and smartphone usage. The goal is to predict student performance in final exams and estimate their final grades. Such a model allows educational institutions and teachers to detect students who may be at risk, enabling them to take early actions to improve academic performance and enhance exam results.

Keywords: Educational Data Mining, Machine Learning, Classification, Student Academic Performance.

I. INTRODUCTION

Education plays a pivotal role in national development, with educational institutions striving to deliver high-quality instruction to students. A critical strategy for achieving this objective involves the early prediction of academic performance, enabling timely interventions to enhance learning outcomes and optimize pedagogical approaches. In recent years, the intersection of Data Mining (DM), Machine Learning (ML), and educational systems has given rise to Educational Data Mining (EDM), an emerging discipline dedicated to extracting actionable insights from academic data to improve teaching and learning processes.

Predicting student performance remains a complex challenge, requiring the integration of computational theory, database management, and ML techniques. The quality of an educational system is intrinsically linked to societal advancement, making it a focal point of contemporary research. By leveraging data from institutional records and educational surveys, ML-based classification methods can significantly enhance academic performance.

Historically, educational institutions accumulate vast datasets encompassing student records, faculty information, and administrative operations. However, much of this data remains underutilized, often limited to basic reporting rather than strategic decision-making. The sheer volume and intricacy of this information further complicate its effective processing, leaving its potential largely untapped. Transforming raw data into meaningful insights is essential for generating knowledge that supports informed institutional policies.

Machine Learning offers a robust framework for identifying patterns within large datasets. As a rapidly evolving field, ML provides diverse analytical techniques applicable across multiple domains, including education. Educational Data Mining (EDM), a specialized branch of ML, focuses on developing novel methodologies to analyze academic data, thereby improving educators' understanding of student learning behaviors. In academia, ML facilitates enhanced educational standards, efficient institutional management, and multi-level assessments of skill development. Given the extensive data stored in educational databases, specialized ML classification techniques have been developed to extract meaningful information, uncover latent relationships, and support data-driven decision-making.

This study examines real-world student data from two primary sources: academic records (containing grades and attendance) and questionnaires (capturing demographic, social, and institutional factors such as age and maternal education level). The analysis aims to predict academic performance while identifying key variables influencing success or failure.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

This study analyzes real-world student data obtained from two primary sources: academic records and survey responses. While the academic records provided limited variables—primarily grades and attendance—the supplementary questionnaire data enriched the dataset by incorporating demographic, social, and institutional factors, such as student age and maternal education level. The primary objectives of this research are (1) to predict academic performance and (2) to identify key determinants influencing student success or failure.

II. RELATED WORK

Xiaofeng Ma and collaborators developed a predictive model employing decision tree algorithms, utilizing a dataset sourced from the UCI Machine Learning Repository. Prior to model construction, the data underwent normalization to ensure consistency. The decision tree was built by selecting the attribute with the highest information gain—calculated via entropy—as the root node, followed by recursive partitioning until all records were classified into distinct leaf nodes. The model's efficacy was subsequently validated using an independent test dataset.

Huda Al-Shehri and her research team introduced two predictive models to forecast students' final examination results, leveraging a publicly available dataset from the University of Minho, Portugal. This dataset comprised 395 student records in mathematics, with the objective of facilitating early academic interventions and informed decision-making. While prior studies predominantly applied the K-Nearest Neighbors (KNN) algorithm—yielding suboptimal accuracy—Al-Shehri et al. implemented both KNN and Support Vector Machine (SVM) for comparative analysis. Their findings indicated marginally superior performance by SVM, evidenced by a correlation coefficient of 0.96 compared to KNN's 0.95.

Pauziah Mohd Arsad and colleagues explored the application of Artificial Neural Networks (ANN) to predict academic performance, measured by Cumulative Grade Point Average (CGPA). The study utilized data from the Electrical Department at Universiti Teknologi MARA, Malaysia, with first-semester grades serving as input variables and eighth-semester CGPA as the target output. The research further compared outcomes between Matriculation and Diploma entry pathways. Model evaluation metrics included the correlation coefficient (R) and Mean Squared Error (MSE), revealing that core subjects in the first and third semesters significantly influenced final CGPA.

In a separate investigation, Kayah et al. employed Naïve Bayes and J48 classification algorithms on a UCI Machine Learning Repository dataset. Using WEKA software, the researchers enhanced model accuracy through discretization, a preprocessing technique that converts continuous variables into categorical intervals.

Method and analysis which is performed in your research work should be written in this section. A simple strategy to follow is to use keywords from your title in first few sentences.

Subheading

Subheading should be Font Size- 10pt, Font Type- Cambria, justified.

Subheading

Subheading should be 10pt Times new Roman,

III. DATASET DESCRIPTION

Data Collection and Preprocessing Methodology

This study employs a dataset gathered from a diverse student population. While government investments in Information Technology have increased, many public higher education institutions continue to rely on manual data management systems. Consequently, the research data were obtained from two complementary sources: institutional academic records and structured student questionnaires.

The institutional records provided limited academic metrics, including final course grades and attendance records (measured by absence frequency). To enrich this dataset, supplementary data were collected through carefully designed Google Forms questionnaires utilizing closed-ended questions.

These instruments captured:

Demographic characteristics (e.g., maternal education attainment, household income level)

Academic history (e.g., number of accumulated backlogs)

Technological access factors (e.g., internet availability, mobile device usage patterns)

The inclusion of both socioeconomic and technological variables enabled comprehensive analysis of potential influences on academic performance outcomes.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

During data preprocessing, categorical variables were transformed into binary numerical representations, with affirmative responses ("yes") coded as 1 and negative responses ("no") as 0. Following this transformation, the finalized dataset contained 30 numerical attributes per student record, with the exception of the final grade variable which retained its original format.

Table 1: Student Dataset Description

S. No	Feature Name	Description
1	**Sex**	Gender of the learner (binary: 'F' – female, 'M' – male)
2	**Student_Age**	Age of the student in years (numeric range: 15–22)
3	**Home_Location**	Type of residence (binary: 'U' – urban, 'R' – rural)
4	**Household_Size**	Number of family members (binary: 'LE3' – ≤3, 'GT3' – >3)
5	**Parent_Marital_Status**	Parents' living arrangement (binary: 'T' – together, 'A' – separated)
6	**Mother_Education_Level**	Mother's highest education level (0 – none, 1 – primary, 2 – middle school, 3 – high school, 4 – college/university)
7	**Mother_Profession**	Mother's occupation (categorical: 'teacher', 'healthcare', 'civil_service', 'homemaker', 'other')
8	**Father_Education_Level**	Father's highest education level (0 – none, 1 – primary, 2 – middle school, 3 – high school, 4 – college/university)
9	**Father_Profession**	Father's occupation (categorical: 'teacher', 'healthcare', 'civil_service', 'homemaker', 'other')
10	**Institution_Choice_Reason**	Reason for selecting the institution (categorical: 'home proximity', 'reputation', 'course', 'other')
11	**Primary_Guardian**	Main guardian (categorical: 'mother', 'father', 'other')
12	**Commute_Time**	Travel time from home to campus (1 – <15 min, 2 – 15–30 min, 3 – 30–60 min, 4 – >1 hr)
13	**Weekly_Study_Hours**	Hours allocated to study per week (1 – <2 hrs, 2 – 2–5 hrs, 3 – 5–10 hrs, 4 – >10 hrs)
14	**Past_Backlogs**	Number of previous subject failures (numeric: 1–3, 4 or more)
15	**Academic_Support**	Access to additional academic support from institution (binary: yes/no)
16	**Parental_Academic_Support**	Educational support provided by family (binary: yes/no)
17	**Private_Tuition**	Enrollment in extra paid tuition classes (binary: yes/no)
18	**Co_Curricular_Activities**	Participation in extracurricular programs (binary: yes/no)
19	**Pre_School_Attendance**	Attended nursery or pre-primary education (binary: yes/no)
20	**Higher_Education_Interest**	Desire to pursue further studies (binary: yes/no)
21	**Home_Internet_Access**	Internet availability at home (binary: yes/no)
22	**Family_Relationship_Quality**	Strength of family relationships (scale: 1 – very poor to 5 – excellent)
23	**After_Class_Free_Time**	Level of leisure time post classes (scale: 1 – very low to 5 – very high)
24	**Social_Activity_Level**	Frequency of social outings with peers (scale: 1 – rarely to 5 – very often)

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

25	**Health_Status**	Overall health condition (scale: 1 – very poor to 5 – very good)
26	**Total_Absences**	Total number of days absent from classes (numeric: 0–93)
27	**Owns_Smartphone**	Whether the student owns a smartphone (binary: yes/no)
28	**Smartphone_Usage_Duration**	Daily mobile usage in hours (binary: 'LE4' – ≤4 hrs, 'GT4' – >4 hrs)
29	**Owns_PC_or_Laptop**	Ownership of a personal computer or laptop (binary: yes/no)
30	**Final_Grade**	Student's academic performance (target variable): 0–3 = Fail, 4–7 = Average, 8–10 = Good

IV. PROPOSED SYSTEM

The classification model employed in this study undergoes a two-phase implementation process: training and testing [12][13]. During the testing phase, the trained classifier utilizes 29 input attributes to predict values for the designated target variable. Predictive accuracy is evaluated distinctly for different prediction types. The methodological framework consists of the following sequential stages:

1. Data Acquisition and Preparation

The analytical pipeline commences with dataset loading into the computational environment, enabling subsequent processing and analysis.

2. Statistical Characterization

Key distribution parameters are computed to facilitate probabilistic modeling:

Mean (μ): Serves as the measure of central tendency, establishing the distribution centroid for normal probability estimation

Standard Deviation (σ): Quantifies the dispersion of attribute values within each class, providing critical information about data variability

These parameters collectively define the expected distribution characteristics for each attribute during probability estimation.

3. Class-Specific Data Partitioning

The dataset is systematically partitioned according to class labels, enabling the development of class-specific statistical profiles.

4. Training Data Summarization

The classification model generates and stores comprehensive summary statistics from the training dataset, including:

Class-specific mean values for each attribute

Corresponding standard deviation measures

These summaries form the probabilistic foundation for subsequent classification decisions.

5. Predictive Inference

The prediction phase employs the normal probability density function:

$P(x|class) = (1/\sqrt{(2\pi\sigma^2)}) * e^{[-(x-\mu)^2/(2\sigma^2)]}$

where:

x = attribute value

μ = class mean

σ = class standard deviation

For each test instance, the model computes attribute probabilities across all classes, ultimately assigning the instance to the class with the highest joint probability.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

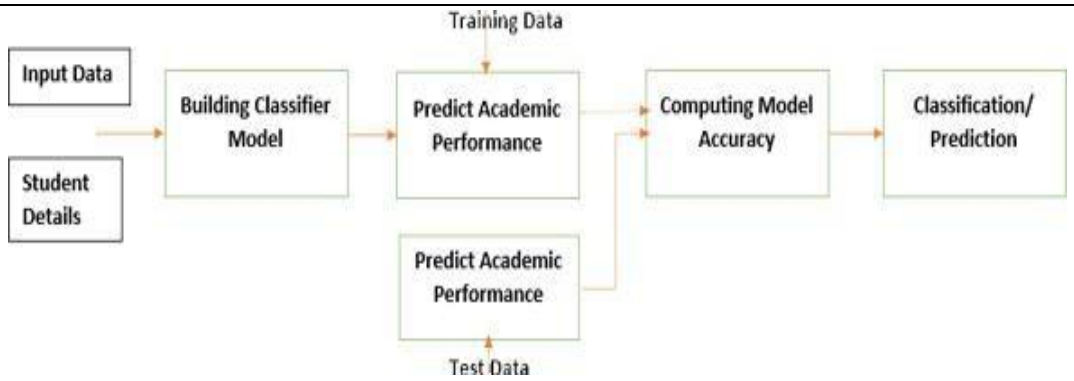

Figure 1: System Architecture.

V. RESULTS

The experimental results reveal significant variation in predictive performance across algorithms. Gradient Boosting emerged as the superior methodology, achieving near-perfect classification (99.66% accuracy), while K-NN demonstrated relatively weaker performance (78.86% accuracy). These findings suggest that ensemble methods, particularly boosting techniques, may offer optimal solutions for educational performance prediction tasks.

The complete comparative results are systematically presented in Table 1, which details the accuracy metrics and error rates for all evaluated algorithms. This quantitative comparison provides valuable insights for researchers and practitioners selecting appropriate machine learning techniques for educational data mining applications

This result investigates the predictive capacity of machine learning algorithms in forecasting student academic performance based on multiple input variables. We employ a comparative analytical framework to evaluate five distinct classification algorithms: K-Nearest Neighbor (K-NN), Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosting (GDBoost). Our methodology focuses on both predictive accuracy and error minimization, with particular attention to algorithm-specific parameter optimization.

The research implements a rigorous experimental design, wherein each algorithm was systematically configured and evaluated:

Decision Tree Algorithm:
Splitting criterion: Gini impurity measure
Branching strategy: Binary partitioning
Depth constraint: Maximum 15 levels
Performance metrics: 93.28% accuracy (6.82% error rate)
Support Vector Machine (SVM):
Kernel optimization: Radial basis function
Margin maximization approach
Performance metrics: 98.60% accuracy (1.40% error rate)
K-Nearest Neighbor (K-NN):
Distance metric: Euclidean
Neighbor parameter: k=5
Performance metrics: 78.86% accuracy (demonstrating comparative limitations)
Random Forest:
Ensemble size: 100 estimators
Depth constraint: Maximum 14 levels (to mitigate overfitting)
Performance metrics: 91.61% accuracy (8.39% error rate)
Gradient Boosting (GDBoost):
Learning rate: 0.1
Iterative optimization: 100 boosting stages
Performance metrics: 99.66% accuracy (0.34% error rate)

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:07/July-2025 Impact Factor- 8.187 www.irjmets.com

Table 2: Accuracy of Models

Classifier	Error Rate	
K-Nearest Neighbour	21.14%	71.86%
Support Vector Machine	1.40%	98.60%
Decision Tree	6.82%	93.28%
Random Forest	8.39%	91.61%
Gradient Boost	0.34%	99.66%

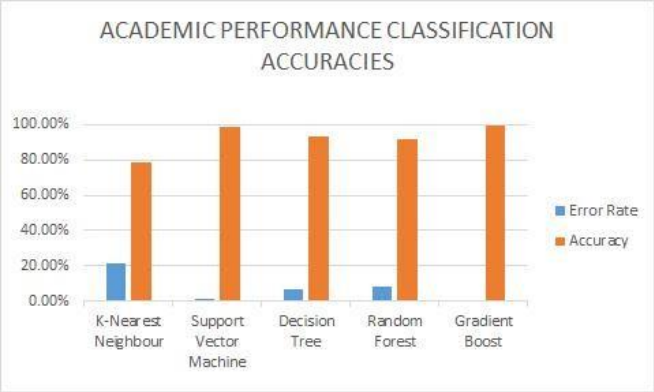

Figure 2: Accuracy of Models

VI. CONCLUSION

Education serves as a fundamental pillar of societal development, necessitating robust methodologies to assess both institutional effectiveness and individual learner outcomes. This study presents an automated predictive system designed to evaluate multiple dimensions of student engagement, including academic performance, social behavior, and technology utilization patterns. While initial validation has confirmed the model's predictive accuracy, further research is required to fully characterize its technological robustness and scalability.

Future investigations will explore two key directions:

The integration of advanced deep learning architectures to enhance predictive capabilities.

A systematic comparative analysis of classification algorithms to identify optimal approaches for different educational assessment scenarios.

This ongoing research aims to establish empirically validated frameworks for educational analytics, contributing to the development of more adaptive and effective learning environments. The findings will provide valuable insights for educators and policymakers seeking to implement data-driven decision-making processes in academic institutions.

VII. REFERENCES

[1] Hashmia Hamsa, Simi Indiradevi, Jubilant J. Kizhakkethottam, Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm, Procedia Technology, Volume 25, 2016, Pages 326-332.

[2] S. S. Athani, S. A. Kodli, M. N. Banavasi and P. G. S. Hiremath, "Student academic performance and social behavior predictor using data mining techniques," 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, 2017, pp. 170-174.

[3] Xiaofeng Ma and Zhurong Zhou. "Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree", 978-1-5386-4649-6/18/$31.00 ©2018 IEEE.

[4] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O. Olatunji, "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor", Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.

[5] Pauziah Mohd Arsad, Norlida Buniyamin, and Jamalul-lail Ab Manan, "A Neural Network Students' Performance Prediction Model (NNSPPM)", IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), 26–27 November 2013.

[6] Kayah, F., "Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers", University of Maryland Publications, College Park, MD, USA.

[7] David, L. M., & Carlos E. G., "Data Mining to Study Academic Performance of Students of a Tertiary Institute", American Journal of Educational Research, vol. 2, no. 9, pp. 713–726, 2014. DOI: 10.12691/education-2-9-3

[8] Romero, C., & Ventura, S., "Educational Data Mining: A Survey from 1995 to 2005", Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, 2007.

[9] Anuradha, C., & Velmurugan, T., "A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance", Indian Journal of Science and Technology, vol. 8, no. 15, 2015. DOI: 10.17485/ijst/2015/v8i15/74555

[10] Cortez, P., & Silva, A., "Using Data Mining to Predict Secondary School Student Performance", In A. Brito & J. Teixeira (Eds.), Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008), pp. 5–12, Porto, Portugal, April 2008. EUROSIS. ISBN: 978-9077381-39-7.

[11] Exploratory Data Analysis, Towards Data Science. Available at: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[12] Chandra, E., & Nandini, K., "Predicting Student Performance Using Classification Techniques", Proceedings of SPIT-IEEE Colloquium and International Conference, Mumbai, India, pp. 83–87.

[13] Huang, S., & Fang, N., "Work in Progress - Prediction of Students' Academic Performance in an Introductory Engineering Course", Proceedings of the 41st ASEE/IEEE Frontiers in Education Conference, 2011, pp. 11–13.