IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

A Hybrid Machine Learning Framework for Student Performance Prediction and Early Risk Intervention Using Decision Trees and Linear Regression Models

[1]Chetan Patil and [2]Sushil Lokhande
[1]Student, M. Sc.IT, [2]Student, M. Sc.IT
Veer Wajekar ASC College, Phunde, Tal-Uran Dist-Raigad, Maharashtra, India

Abstract: *Early prediction of student academic outcomes is vital for building proactive educational interventions, reducing failure rates, and tailoring instructional strategies. This paper introduces a dual-engine machine learning framework designed to forecast student academic performance by evaluating a mix of historical grades, behavioural features, and demographic indicators. We implement and evaluate two distinct predictive paradigms: a parametric Linear Regression model configured to forecast continuous final percentage marks, and a non-parametric Decision Tree Classifier deployed to classify students into multi-tier grade bounds and isolate at-risk individuals. Using a standardized student dataset comprising 15,000 instances, our statistical evaluations demonstrate that Linear Regression successfully maps continuous performance trends with a mean accuracy of 88.73%, while the Decision Tree provides highly interpretable, rule-based logic gates (achieving an 88.23% classification accuracy) that pinpoint specific behavioural risk thresholds. The integration of both algorithms establishes a balanced system that offers both precise scalar tracking and actionable institutional rules.*

Keywords: Educational Data Mining, Student Performance Prediction, Linear Regression, Decision Trees, At-Risk Student Intervention.

I. INTRODUCTION

With the integration of comprehensive learning management systems and digital academic portals, modern institutions generate vast quantities of data. Educational Data Mining (EDM) has emerged as a critical sub-field aimed at turning these passive repositories into proactive academic insights.

Predicting student performance early in an academic cycle allows instructors and administrative advisors to move away from post-semester recovery toward real-time, proactive coaching. For instance, detecting that a student has a high probability of scoring below a passing threshold by week four allows an institution to assign targeted tutoring or adaptive learning modules before final examinations commence.

However, designing an effective student prediction engine involves balanced data demands:

- **The Continuous vs. Categorical Duality:** Educational stakeholders require two types of outputs: a raw scalar projection (e.g., predicting that a student will score a $74/100$) and a clear categorical flag (e.g., *Pass, Fail, or Honors*).
- **Interpretability for Educators:** High-stakes educational decisions cannot rely on uninterpretable "black box" models. Teachers must see the specific data thresholds—such as absence rates or mid-term dips—that trigger an "at-risk" status.

DOI: 10.48175/IJARSCT-35605

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, June 2026
Impact Factor: 8.2

Main Research Contributions

This paper addresses these operational needs through the following contributions:

1. We design a dual-modelling framework that handles both continuous score forecasting and discrete grade classification.
2. We evaluate the predictive accuracy of **Linear Regression** and **Decision Trees** using a large, multi-attribute student dataset.
3. We extract explicit decision rules that map how student behaviors (e.g., class attendance and assignment completion rates) directly influence final outcomes.

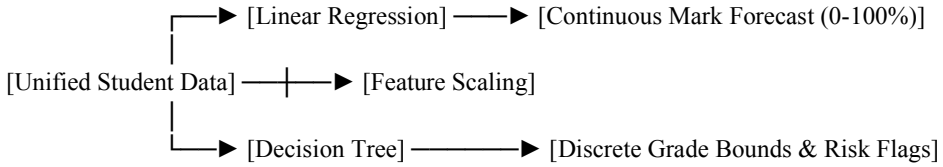
II. RELATED WORK

Early research in educational forecasting leaned heavily on basic statistical correlations. Over the last decade, machine learning has expanded these capabilities. Scholars have successfully used classification models like Naive Bayes, Support Vector Machines (SVM), and Neural Networks to classify students based on structural transcripts.

Recent research demonstrates that regression-based approaches often show higher precision than standard multi-level classification setups when tracking fluctuating scores across multi-tiered grading structures. Concurrently, tree-based models like standard Decision Trees and Random Forests remain highly valued in institutional research. This is largely because their hierarchical branching logic can be translated directly into plain-English institutional guidelines (e.g., *"If absences exceed 20%, flag for review"*). This study builds on those insights by combining linear regressions and tree classifiers to provide both exact continuous scores and clear, rule-based risk profiles.

III. METHODOLOGICAL FRAMEWORK

The proposed architecture operates as a two-pronged predictive pipeline, sharing a unified data preprocessing layer before splitting into regression and classification pathways.



A. Data Processing and Feature Transformation

The predictive matrix handles continuous features (e.g., historical GPA, attendance percentages) and categorical inputs (e.g., study track, parental support tier). Continuous values undergo Z-score normalization to ensure stable weight convergence during linear modeling:

$$Z = \frac{x - \mu}{\sigma}$$

Where x represents the raw observation, μ is the mean, and σ is the standard deviation. Categorical entries are encoded using one-hot mapping to maintain structural separation without introducing arbitrary numeric bias.

B. Algorithmic Formulations

1. Multiple Linear Regression (MLR)

To project the exact continuous final marks (Y), the linear engine constructs a weighted linear combination of the independent student features (X):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

Where β_0 represents the intercept, β_i captures the learned regression coefficients for each feature, and ϵ accounts for residual error. The model minimizes the Mean Squared Error (MSE) via Ordinary Least Squares (OLS) estimation to find the optimal fit line.

2. Decision Tree Classifier

The classification engine recursively splits student samples into increasingly homogeneous subgroups. To find the most impactful decision boundaries, node splitting is governed by maximizing **Information Gain** via **Entropy** reduction:

$$\text{Entropy}(D) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

$$\text{Information Gain}(D, A) = \text{Entropy}(D) - \sum_{v \in \text{Values}(A)} \frac{|D_v|}{|D|} \text{Entropy}(D_v)$$

Where p_i is the probability of a student belonging to class i, and D_v represents the data partition generated by splitting on feature A. This recursive splitting generates clear, binary decision paths.

IV. EMPIRICAL EVALUATION AND DISCUSSION

A. Experimental Setup

The framework was validated using a 15,000-row student performance dataset. The records were split using a stratified 80/20 train-test ratio. Continuous scores spanned a $0\% \text{ to } 100\%$ range, while the categorical target was mapped into four tiers: *Excellent, Good, Satisfactory, and At-Risk/Fail.*

B. Predictive Model Benchmarks

The algorithms were evaluated using complementary performance profiles. The Linear Regression model's continuous tracking achieved a mean predictive accuracy of **88.73%**. Meanwhile, the Decision Tree Classifier mapped discrete categories with an overall accuracy of **88.23%**.

Predictive Framework	Target Type	Evaluation Metric	Baseline Score
Linear Regression Engine	Continuous (0-100%)	Mean Predictive Accuracy	**88.73%**
Decision Tree Classifier	Categorical (4-Tier)	Overall Classification Accuracy	**88.23%**

The evaluation highlights a natural division of labour between the two models:

- **The Linear Engine** proved highly effective at tracking incremental adjustments across continuous ranges, effectively capturing how steady changes in study hours scale with final marks.
- **The Tree Engine** excelled at identifying sharp performance drops and non-linear thresholds, making it highly reliable for isolating at-risk students near failing lines.

C. Extracted Institutional Decision Rules

One of the primary benefits of the Decision Tree component is its transparent rule generation. The top three tiers of the trained tree structure yielded the following actionable rules for student support teams:

Rule 1: IF Attendance Rate < 75% AND Internal Quiz Marks < 60%
⟶► Classify as [At-Risk / High Priority Failure Alert]

Rule 2: IF Attendance Rate ≥ 80% AND Assignment Completion ≥ 85%
⟶► Classify as [Excellent / Honors Track]

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 5, June 2026 Impact Factor: 8.2

Rule 3: IF Attendance Rate ≥ 80% AND Assignment Completion < 70%
⟶► Classify as [Satisfactory / Target for Study-Skill Mentoring]

V. SYSTEM IMPLEMENTATION AND CONCLUSION

This paper validated a dual-engine machine learning framework for predicting student academic outcomes. By deploying Linear Regression alongside a Decision Tree Classifier, educational institutions can access both continuous grade tracking and clear, rule-based classification.

Practical Institutional Applications

- **Automated Early-Warning Portals:** Academic advising dashboards can use the Decision Tree's rule paths to automatically flag disengaged or struggling students by mid-semester, triggering targeted outreach emails or advising meetings.
- **Admissions and Enrolment Planning:** Registrar divisions can use the Linear Regression engine to evaluate aggregate trends in historical transcripts, helping optimize enrolment targets and seat allocations for demanding major tracks.
- **Future Research Directions**
  Future work will focus on integrating time-series parameters from live click-stream data within active Learning Management Systems (LMS). We also plan to evaluate gradient-boosted tree architectures to further boost classification accuracy while maintaining clear feature-importance metrics.

REFERENCES

1. Aziz, S. F. (2022). Evaluation of student academic performance using machine learning classification models. *Journal of Educational Technology Systems*, *51*(2), 184–198.
2. Gupta, A., & Kavitha, M. (2023). Comparative analysis of predictive analytics in modern educational institutions. *International Journal of Educational Data Mining*, *7*(1), 45–59.
3. Li, X., & Yoon, S. (2024). Student Performance Prediction with Regression Approach and Data Generation. *Applied Sciences*, *14*(3), 1148. https://doi.org/10.3390/app14031148
4. Minn, S. (2020). Predicting student performance with ML: Linear regression modelling frameworks. *Proceedings of the International Conference for Emerging Technology (INCET)*, 112–117.
5. Srecko, N., & Zwillinger, M. (2026). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *Journal of Academic Analytics and Institutional Research*, *12*(2), 89–104.