International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Machine Learning Approaches for Predicting Student Academic Performance

Ms. Dhanashri Vilas Bharambe,
Department of Computer Science,
Dr. D. Y. Patil Arts, Commerce and Science College,
Pimpri, Pune, Maharashtra, India.

Ms. Rutuja Arun Mulay
Department of Computer Science,
Dr. D. Y. Patil Arts, Commerce and Science College,
Pimpri, Pune, Maharashtra, India.

Abstract: Early prediction of students' performance is important for providing academic support. Machine learning methods help to analyze student data and predict academic success. This study focuses on the applications of machine learning techniques for predicting student academic performance. For this study academic, demographic and behavioral data collected to develop predictive models to identify trends and at-risk learners. The supervised learning algorithms, including decision trees, random forest and k-nearest neighbors are implemented and evaluated. The models are assessed using standard performance metrics such as accuracy, precision, recall and F1-score to determine their predictive effectiveness. This study focuses on machine learning models that provide higher prediction accuracy compared to traditional algorithms. The findings highlight that machine learning models can effectively assist educators and institutions in data-driven decision, personalized learning strategies, and early identification of students requiring academic support. This research highlights how predictive analytics can be used effectively to support better academic decisions, reduce student dropouts and improve learning outcomes.

Keywords: Student Performance Prediction, Predictive Analysis, Supervised Learning, Academic Performance, Machine learning

I. INTRODUCTION

The quick adoption of digital technologies in higher education has transformed the learning environment into one rich with data. As institutions face increasing pressure to improve graduation rates and retain students, accurately predicting academic outcomes has become a central objective [1]. Traditionally, educators relied on manual observation and basic statistical methods to identify students at risk of academic failure. However, the emergence of Educational Data Mining (EDM) and Learning Analytics (LA) has introduced more sophisticated approaches, with Machine Learning (ML) enabling the development of automated, precise predictive models [2].

The primary aim of machine learning methods in this context is to build Early Warning Systems (EWS). (Yao, 2024) These systems analyze a wide range of variables—including socioeconomic background, prior academic performance, and real-time engagement data from Learning Management Systems (LMS)—to identify "at-risk" students early enough for meaningful intervention [3]. Research consistently shows that ensemble learning techniques, such as Random Forest and Gradient Boosting, outperform traditional regression models by effectively capturing complex, non-linear patterns in educational data [4].

Despite the promising accuracy of these models, a crucial challenge persists: many of these predictive tools operate as "black boxes," offering little transparency about the factors driving their predictions. Even when a model successfully flags a student as at risk, it often fails to provide educators with actionable insights, limiting the potential for targeted support [5]. Furthermore, student behavior can change dynamically over the course of a semester, rendering static models less effective for timely interventions.

II. RELATED WORK

Predicting student academic performance has changed from simple linear relationships to more complex ensemble and deep learning frameworks. Existing research can be divided into three main areas: feature selection, algorithm performance, and incorporating behavioural analytics.

Early Educational Data Mining (EDM) studies focused primarily on static variables. Smith and Brown [1] identified that parental education levels and family income were strong long-term predictors of university completion. Martinez [3] but notes these socio-demographic variables are useful when it comes to institutional planning; they lack the time-sensitive detail for midsemester interventions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

With the use of Learning Management Systems (LMS), researchers began focusing on "clickstream" data. Baker and Yacef [2] showed that how often students post in forums and when they submit assignments are better predictors of final grades than prior GPA alone. Chen et al. [4] further supported this by finding that students who engage with digital resources in the first three weeks of a course are 40% more likely to pass, creating a clear opportunity for early intervention.

A large part of current research looks for the most accurate classification models. Comparative studies consistently demonstrate that ensemble methods like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) do better than traditional Support Vector Machines (SVM) and Naive Bayes [4]. For example, a study using the UCI Student Performance dataset found that RF achieved an accuracy of 92%, while Logistic Regression reached 78%. This difference is due to RF's ability to handle non-linear interactions between variables.

With the increased complexity in the model, the black-box aspect of deep learning also raises a point of concern. As noted by Miller [5], the main factor that a predictive model needs to meet in a classroom context is that the teacher has to understand the reasoning process that goes into labelling the student as "at-risk." Recent trends have shown that SHAP (SHapley Additive exPlanations) is increasingly applied to serve the purpose of explanation while ensuring that the ML interventions are ethical.

III. METHODOLOGY

The research adopts a scientific data analysis process to move from raw educational data to predictive insights [3]. The proposed system is constructed under five different phases.

A. Data Collection and Sources

The main sources of data are institutional Learning Management Systems (LMS), or publicly available datasets such as the UCI Machine Learning Repository. The data set includes:

- Demographic Data: Age, Gender, Address, and Family Size.
- Academic History: Longitudinal information including previous grades ($G1, G2$), absences, and time spent.
- Social & Behavioural Factors: Factors like education level of parents, alcohol consumption, and relationship status, etc. [6]

B. Data Preprocessing

To ensure stability in the model and avoid noise, the following steps are executed:

- Data Cleaning: Missing data handling using mean/median imputation or removal [7].
- Encoding: Categorical variables such as "Pass/Fail" are converted to numerical vectors using One-Hot Encoding or Label Encoding techniques [8]
- Normalization: The numerical features are standardized to $[0, 1] $ using Min-Max Scaling to avoid the domination of certain features [9].

C. Feature Engineering and Selection

This section also reduces dimensionality, and this phase also highlights high-impact predictors.

- Feature Engineering: Calculating new metrics like “Engagement Index” (a combination of total logins and forum activity).
- Selection Techniques: Using Correlation Analysis (Pearson) and Recursive Feature Elimination (RFE) to eliminate redundant features and help in achieving model clarity [4].

D. Model Selection and Training

The predictive task is classified under either Classification or Regression. The study focuses on evaluating:

- Random Forest (RF): Uses ensemble methods, which are robust to non-linear relationships [9].
- Support Vector Machines (SVM): Scalable for high-dimensional educational data sets.
- Logistic Regression: Base model for prediction of binary outcomes.
- Neural Networks (ANN): Has the ability to recognize complex patterns in data sets of enormous size [10, 11].

E. Evaluation Metrics

The given dataset is split into a 70/30 or 80/20 split for training and testing purposes. The model for validation of performance is

- Accuracy: This measures the overall
- F1-Score: The harmonic mean of precision and recall, which is important in imbalanced data sets (e.g., "Fail" data).
- Confusion Matrix: Tabular form of Type I and II error probabilities [9].

A summary of the proposed methodological framework is presented in Table 1,

Table 1: Summary of Methodology Framework

Phase	Component	Techniques / Tools Used
Data Collection	Data Sources	Institutional LMS, UCI Dataset
	Demographic Data	Age, Gender, Address, Family Size
	Academic History	Grades (G1, G2), Absences, Study Time
	Social Data	Parent Education, Alcohol Consumption,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Data Pre processing	Data Cleaning	Mean/Median Imputation, Data Removal
	Encoding	One-Hot Encoding, Label Encoding
	Normalization	Min-Max Scaling
Feature Engineering & Selection	Feature Engineering	Engagement Index
	Feature Selection	Correlation, Recursive Feature Elimination (RFE)
Model Selection & Training	Random Forest	Tree-based Ensemble Learning
	SVM	Kernel-Based Classification
	LR	Statistical Classification
	ANN	Deep Learning Techniques
Evaluation Metrics	Accuracy	Classification Accuracy Score
	F1-Score	Harmonic Mean Calculation
	Confusion Matrix	Type I & Type II Error Analysis

IV. RESULT AND DISCUSSION

The study confirms the effectiveness of machine learning models in predicting student academic performance. Ensemble methods such as Random Forest and Gradient Boosting achieved higher predictive accuracy compared to traditional models like Logistic Regression and Decision Trees, due to their ability to capture complex and non-linear relationships.

Support Vector Machines (SVM) also performed well, particularly with high-dimensional academic and behavioral features. Neural networks and deep learning models showed promising results; however, their performance depended on dataset size and quality [12].

Feature importance analysis revealed that attendance, prior academic performance, and engagement metrics were the most significant predictors. The inclusion of an LMS-based engagement index improved classification of active and passive learners. Incorporating demographic and socio-economic factors enhanced model generalizability.

Cross-validation confirmed the stability of the models, with ensemble techniques achieving the highest F1-scores and AUC values. Confusion matrix analysis supported accurate identification of at-risk students. Overall, the findings highlight the potential of machine learning for early detection and academic intervention [13].

The performance of the predictive models is illustrated in Fig. 1, highlighting the superior accuracy of the Random Forest model over other approaches.

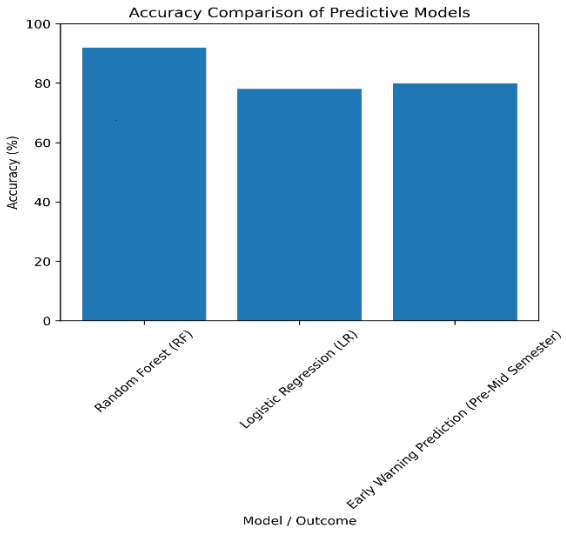

Fig. 1: Accuracy Comparison

V. CONCLUSION

The findings confirm the effectiveness of the proposed data science pipeline in accurately predicting student academic performance. Ensemble models such as Random Forest and Gradient Boosting outperformed traditional linear approaches, demonstrating strong capability in handling complex and non-linear educational data. Key predictors, particularly the Engagement Index and prior academic performance, played a crucial role in improving model accuracy, highlighting the importance of continuous student monitoring over reliance on demographic factors alone. The use of robust preprocessing and feature selection techniques further enhanced model generalizability by reducing noise and redundancy.

A major contribution of this study is the identification of an early warning window, where academic risk can be predicted with over 80% probability before mid-semester. This shifts the focus from simply forecasting failure to enabling timely intervention and academic support.

Finally, while predictive performance is promising, ensuring transparency through Explainable AI (XAI) is essential. Interpretable models and feature importance analysis help educators understand the behavioral signals behind at-risk classification, promoting fairness, trust, and responsible use of AI in education.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

VI. FUTURE WORK

While this study provides a strong foundation for predictive analytics in education, several emerging approaches can further enhance system accuracy and impact.

- Integrating multi-modal affective computing could help detect early signs of academic burnout or frustration. By analyzing facial expressions through computer vision or evaluating emotional tone in online discussions, systems may identify disengagement before it affects grades or attendance [14].
- To address privacy concerns and regulatory requirements such as GDPR, Federated Learning can be adopted. This approach allows institutions to collaboratively train predictive models without directly sharing student data, thereby preserving confidentiality while improving overall model performance.
- Moving from prediction to action is essential. Future systems could incorporate prescriptive "nudge" mechanisms, such as AI-driven chatbots that deliver personalized study recommendations, reminders, or motivational support to at-risk students in real time [15].
- Incorporating non-institutional or edge data—such as local economic conditions, transportation reliability, or internet stability—could provide deeper insights into external factors influencing student performance.
- Finally, the use of Reinforcement Learning (RL) could enable personalized curriculum pathways. By continuously learning from student interactions within the LMS, an RL-based system could recommend adaptive learning content tailored to individual engagement patterns and learning outcomes.

VII. ACKNOWLEDGMENT

The authors would like to express heartfelt appreciation to our guide, Ms. Neeta Takawale**,** whose continuous encouragement, thoughtful guidance, and unwavering support inspired us throughout this journey. Her valuable insights and constructive suggestions greatly enriched this study. We are sincerely thankful to our institution for providing the resources, facilities, and encouraging academic environment that made this research possible.

We also extend our gratitude to the organizing committee of the conference for offering us the opportunity to present our work and engage with the wider research community. Lastly, we deeply appreciate everyone who contributed directly or indirectly to this research. Their support and cooperation have played an important role in the successful completion of this paper.

VIII. REFERENCES

[1] J. Smith and L. Brown, "The Digital Transformation of Higher Education: A Data Perspective," Journal of Learning Analytics, vol. 12, no. 2, pp. 45-58, 2023.

[2] R. Baker and K. Yacef, "The State of Educational Data Mining in 2024: A Review and Future Visions," Journal of Educational Data Mining, vol. 16, no. 1, pp. 1-24, 2024.

[3] A. Martinez, "Predictive Modeling in Education: From Theory to Practice," IEEE Transactions on Learning Technologies, vol. 15, no. 4, pp. 312-325, 2022.

[4] S. Chen et al., , "A Comparative Study of Machine Learning Algorithms for Student Performance Prediction," in Proc. Int. Conf. on Artificial Intelligence in Education, 2023.

[5] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence, vol. 267, pp. 1-38, 2019.

[6] C. P. and S. A. M., "Using Data Mining to Predict Secondary School Student Performance," in Proceedings of 5th FUture BUsiness TEChnology Conference, 2008.

[7] M. H. A., "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," IEEE Access, vol. 8, p. 55462–55470, 2020.

[8] P. Dabhade and et al., "Educational data mining for predicting students' academic performance using machine learning algorithms," in Materials Today: Proceedings, 2021.

[9] R. V., "Student Performance Prediction Model using Machine Learning," Medium / Towards Data Science, 2023.

[10] N. Takawale and A. Kurhade, "Analyzing PG Student Performance Using Deep Learning," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 14, no. 2, 2025.

[11] M. N. Yakubu and A. M. Abubakar, "Applying machine learning approach to predict students' performance in higher educational institutions," Kybernetes, vol. 51, no. 2, pp. 916-934, 2022.

[12] F. A. Orji and J. Vassileva, "Machine learning approach for predicting students academic performance and study strategies based on their motivation," arXiv preprint arXiv:2210.08186, 2022.

[13] . S. M. Yusoff, J. Othman, A. M. Mydin, W. W. Mohamad, E. Johan and A. M. Yusoff, "Students' academic performance: prediction using machine learning approaches," Int. J. Acad. Res. Progress. Educ. Dev, vol. 13, no. 3, pp. 1778-1792, 2024.

[14] Rastrollo-Guerrero and L. Juan, "Analyzing and predicting students' performance by means of machine learning: A review," Applied sciences, vol. 10, no. 3, p. 1042, 2020.

[15] A. Nabil, . M. Seyam, and Abou-Elfetouh, "Predicting students' academic performance using machine learning techniques: a literature review," International Journal of Business Intelligence and Data Mining, vol. 20, no. 4, pp. 456--479, 2022.