The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

19. Machine Learning Approach for Predicting Student Academic Performance

Ahaan Pal
BCA Student
Department of BCA, IIMT College Of Management, Gr Noida
Aditya Choudhary
BCA Student
Department of BCA, IIMT College Of Management, Gr Noida

Abstract

This paper presents a comprehensive study on predicting student academic performance using machine learning techniques. The research addresses the significant problem of early identification of academically at-risk students, which enables timely academic interventions and improves overall student outcomes. The study employed multiple supervised learning algorithms including Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting on a dataset of 500 students with 15 academic and behavioral features spanning multiple semesters. The methodology utilized an 80-20 train-test split with stratified cross-validation for robust model evaluation. Results demonstrated that the Random Forest algorithm achieved the highest accuracy of 94.2% in predicting whether students would achieve above-average or below- average grades, with a precision of 92.8% and recall of 95.6%. The study found that attendance rate, previous semester GPA, and class participation were the three most influential features in predicting academic performance, collectively accounting for over 75% of the model's decision- making process. These findings suggest that machine learning models can serve as effective and reliable tools for educational institutions to identify struggling students early and provide appropriate support mechanisms, ultimately leading to improved retention rates and academic success.

Keywords: Machine Learning, Academic Performance, Prediction, Classification, Student Success, Decision Tree, Random Forest, Educational Data Mining, Supervised Learning, Early Warning Systems

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

1. Introduction

1.1 Background of the Study

In recent years, educational institutions have accumulated vast amounts of data related to student performance, attendance, engagement, and behavioral patterns. This explosion of educational data has opened new opportunities for applying data mining and machine learning techniques to extract meaningful insights and patterns that can significantly improve educational outcomes. Traditional methods of identifying struggling students have relied primarily on periodic examinations and subjective evaluations by instructors, which are inherently limited in their scope and timeliness. However, these conventional methods are reactive in nature and often fail to provide early warnings about potential academic difficulties until it is too late for meaningful interventions.

The advent of machine learning technologies has enabled the development of sophisticated predictive models that can identify at-risk students well in advance, allowing educators to implement timely interventions and support mechanisms before academic performance deteriorates further. Machine learning algorithms can process large volumes of multidimensional data and uncover complex, non-linear relationships between various factors and academic outcomes, relationships that might be invisible to human analysts. By leveraging these advanced computational techniques, educational institutions can transition from a reactive approach to a proactive one, focusing on prevention rather than remediation.

1.2 Importance and Motivation of the Topic

Predicting student academic performance is highly significant for several compelling reasons. First and foremost, it enables educational institutions to allocate limited resources more efficiently by identifying students who require additional support and attention. This targeted resource allocation is far more cost-effective than providing support to all students indiscriminately. Second, early detection of at-risk students allows for the implementation of timely academic counseling, mentorship programs, and supplementary instruction before students fall significantly behind. This early intervention can make the difference between academic success and failure, particularly for first-generation and disadvantaged students.

Third, understanding the specific factors that influence academic success can help institutions improve their curricula, teaching methodologies, and support services. Fourth, such predictive systems can significantly boost student retention rates and overall institutional effectiveness in

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

fostering student success, leading to improved graduation rates and better student outcomes. Finally, from a student perspective, early identification and support can reduce stress and anxiety, improve self-efficacy, and increase student engagement and motivation. The motivation for this research stems from the observation that many students struggle academically despite having the potential to succeed, and much of this struggle could be prevented with proper early intervention.

1.3 Research Problem, Questions, and Objectives

The primary research problem addressed in this study is: How can machine learning algorithms be effectively utilized to predict student academic performance with high accuracy and reliability? More specifically, this research seeks to answer the following research questions:

- Which machine learning algorithms provide the best predictive performance for student academic outcomes?
- What are the most influential factors that affect student academic performance?
- How do different machine learning models compare in terms of accuracy, precision, recall, and F1-score?
- Can predictive models generalize across different student populations and institutional contexts?

The specific research objectives of this study are as follows:

- To develop and evaluate multiple machine learning models for predicting student academic performance
- To identify and rank the most influential factors affecting student academic success
- To compare the performance of different machine learning algorithms for this classification task
- To develop practical recommendations for educational institutions based on the findings
- To discuss limitations and provide directions for future research

2. Literature Review

Extensive research has been conducted over the past two decades on predicting student academic performance using various computational approaches. Early studies focused on statistical methods and regression analysis, but recent advances in machine learning have demonstrated significantly superior predictive capabilities. This section reviews the key areas of relevant literature.

2.1 Educational Data Mining and Learning Analytics

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com Email: asianthinkerjournal@gmail.com

Educational Data Mining (EDM) is an emerging discipline that applies data mining techniques specifically to educational data to extract meaningful patterns and insights. It is closely related to the broader field of Learning Analytics, which focuses on collecting, analyzing, and reporting data about learners and their learning contexts. Research in EDM has shown that data mining can effectively identify patterns in student learning behaviors and predict academic outcomes with reasonable to high accuracy. Studies have demonstrated that features such as attendance records, assignment completion rates, previous academic records, and engagement metrics are significant predictors of academic success.

2.2 Machine Learning Classification Algorithms

Several machine learning algorithms have proven particularly effective for classification tasks in educational settings. Decision Trees are popular because they provide interpretability, making them valuable for understanding which specific factors most influence academic performance. Decision trees recursively partition the feature space based on the most informative splits, creating a transparent model that can be easily explained to non-technical stakeholders.

Random Forests address the overfitting problem inherent in single decision trees by building multiple trees on bootstrapped samples and averaging their predictions. This ensemble approach has shown superior performance in many studies and provides feature importance rankings. Support Vector Machines excel at finding optimal decision boundaries through kernel transformations and are particularly effective in high-dimensional feature spaces. Gradient Boosting techniques iteratively improve model predictions by focusing on misclassified examples and often achieve state-of-the-art results. k-Nearest Neighbors is a simple yet effective algorithm that classifies instances based on the majority class of their k nearest neighbors.

2.3 Key Features Influencing Academic Performance

Literature consistently identifies several critical factors affecting student academic performance. Attendance rate is one of the strongest predictors, with research showing that students who attend classes regularly tend to perform significantly better. This relationship is causal in nature: attendance enables students to receive instruction, participate in discussions, and benefit from immediate feedback. Prior academic achievement, represented by previous semester GPA or standardized test scores, is another powerful predictor, reflecting accumulated knowledge and established study habits.

Class participation and engagement metrics indicate student involvement and motivation, both essential for academic success. Study hours reflect student effort and commitment to learning. Assignment completion rates demonstrate work ethic and task completion capability. Quiz and exam performance directly measure knowledge acquisition. Demographic factors such as age,

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com Email: asianthinkerjournal@gmail.com

gender, and socioeconomic status have been found to influence academic outcomes, though their effects are often mediated by other factors. Additionally, psychological variables such as self-efficacy, motivation, and learning style preferences significantly impact academic performance, though these are more difficult to measure quantitatively.

3. Methodology

3.1 Research Design and Approach

This research employs a quantitative approach using supervised machine learning classification techniques. The study follows a systematic empirical methodology in which predictive models are developed and evaluated using historical student data. The research design includes data collection, preprocessing, feature selection, model training, evaluation, and analysis of results. This structured approach ensures rigorous and reproducible results.

3.2 Data Collection and Dataset Description

The study utilized a comprehensive dataset comprising 500 students collected from a higher education institution over a three-year period. This dataset includes 15 carefully selected features representing academic performance, behavioral patterns, and demographic information. The dataset represents a balanced mix of students with varying levels of academic achievement, ensuring representative learning samples for model training.

Features are organized into three categories as presented in Table 3.1 below:

Feature Category	Specific Features
Academic Features	Previous Semester GPA, Quiz Average Scores, Midterm Exam Score, Assignment Completion Rate, Lab Work Scores
Behavioral Features	Attendance Rate (%), Class Participation Score, Weekly Study Hours, Library Usage Frequency, Assignment Submission Timeliness
Demographic Features	Student Age, Gender, First-Generation Student Status, Socioeconomic Status, Scholarship Recipient Status

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

The target variable for this classification task is binary: whether students achieved above-average performance (GPA ≥ 3.0) or below-average performance (GPA < 3.0). This binary classification threshold was chosen as it represents the conventional standard for 'good

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com Email: asianthinkerjournal@gmail.com

academic standing' at most higher education institutions.

3.3 Data Preprocessing and Feature Engineering

Data preprocessing is a critical step that significantly impacts model performance. The preprocessing pipeline involved several sequential steps: First, missing values were handled using appropriate imputation techniques. For numerical features, mean imputation was applied, while for categorical features, mode imputation was used. After imputation, features were normalized using z-score normalization to ensure equal contribution regardless of their original scales. Outliers were identified using the Interquartile Range (IQR) method and removed to prevent their disproportionate influence on model training.

Class imbalance was addressed using oversampling techniques to ensure adequate representation of both classes during training. Exploratory data analysis was performed to understand feature distributions, identify correlations, and detect potential multicollinearity. Feature scaling was applied to all features to normalize their ranges. Finally, categorical variables were encoded using one-hot encoding to convert them into numerical representations suitable for machine learning algorithms.

3.4 Machine Learning Models Implemented

Four machine learning algorithms were carefully selected and implemented based on their proven effectiveness in classification tasks:

- Decision Tree: A tree-based classifier that recursively partitions the feature space by selecting the most informative splits, providing high interpretability
- Random Forest: An ensemble method combining multiple independently trained decision trees, reducing overfitting through bootstrap aggregating
- Support Vector Machine (SVM): A boundary-based classifier that finds optimal separating hyperplanes in potentially transformed feature spaces using kernel methods
- Gradient Boosting: An ensemble technique that sequentially improves predictions by focusing on previously misclassified examples, often achieving superior performance

3.5 Model Evaluation Methodology

Comprehensive evaluation was conducted using multiple performance metrics to assess different aspects of model quality. The dataset was split into 80% training data and 20% testing data, with stratification to ensure both sets have similar class distributions. The primary evaluation metrics include:

- Accuracy: The proportion of correct predictions among all predictions
- Precision: The proportion of true positive predictions among all positive

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

predictions

- Recall: The proportion of true positives correctly identified among all actual positives
- F1-Score: The harmonic mean of precision and recall, providing a balanced performance measure
- ROC-AUC: Area under the Receiver Operating Characteristic curve, measuring performance across all classification thresholds

Additionally, 5-fold stratified cross-validation was performed to ensure robust evaluation and reduce variance in performance estimates. This technique divides the data into five equal-sized folds, trains the model on four folds, and evaluates it on the remaining fold, repeating this process five times. The final performance metric is the average across all five iterations, providing a more reliable estimate of true model performance.

4. Results and Analysis

4.1 Comparative Model Performance Results

Table 4.1 presents comprehensive performance metrics for all four machine learning models evaluated in this study:

Algorithm	Accuracy	Precision	Recall	F1-Score
Decision Tree	85.6%	83.2%	88.1%	0.855
Random Forest	94.2%	92.8%	95.6%	0.942
SVM	89.4%	87.6%	91.2%	0.894
Gradient Boosting	92.8%	91.4%	94.1%	0.928

The results clearly demonstrate that the Random Forest algorithm achieved the highest overall performance with an accuracy of 94.2%, precision of 92.8%, and recall of 95.6%. These metrics indicate that the Random Forest model correctly predicts academic performance for 94.2% of all students, with high reliability in both identifying high-performing students and at-risk students. The high recall score suggests that the model is particularly effective at identifying at-risk students who need intervention.

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

Feature	Importance
Attendance Rate (%)	0.285
Previous Semester GPA	0.268
Class Participation Score	0.198
Assignment Completion Rate	0.156
Weekly Study Hours	0.093

4.2 Comprehensive Discussion of Results

The empirical results strongly support the effectiveness of machine learning approaches for predicting student academic performance. The Random Forest algorithm's achievement of 94.2% accuracy represents a substantial improvement over traditional statistical methods and demonstrates the value of ensemble learning approaches in this domain. The superior performance of Random Forest compared to individual decision trees (85.6% accuracy) clearly illustrates the power of ensemble methods in reducing overfitting and improving generalization to new data.

The feature importance analysis reveals several critical insights about factors influencing student success. Attendance rate emerges as the single most influential factor, accounting for 28.5% of the model's predictive power. This finding has strong theoretical backing in educational research, as regular attendance enables students to receive instruction, participate in discussions, form relationships with peers and instructors, and maintain motivation and engagement. Previous semester GPA is the second most important factor (26.8%), reflecting the reality that academic performance exhibits substantial continuity over time, with students' established study habits and knowledge base creating momentum.

Class participation (19.8%) ranks third, highlighting the significant impact of active engagement in the learning process. Students who actively participate in class discussions and activities demonstrate higher levels of engagement, motivation, and understanding. Together, these three top factors account for over 75% of the model's decision-making process, suggesting that institutions should prioritize monitoring and supporting these three dimensions of student performance.

5. Conclusion and Recommendations

5.1 Summary of Key Findings

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

This comprehensive research successfully developed and evaluated machine learning models for predicting student academic performance, achieving an accuracy of 94.2% with the Random Forest algorithm. The study identified attendance rate, previous semester GPA, and class participation as the three most influential factors affecting academic success. These findings confirm that machine learning approaches can provide educational institutions with powerful tools for early identification of at-risk students, enabling timely and targeted interventions.

5.2 Study Limitations

While this research presents significant findings, several limitations should be acknowledged. First, the dataset comprised students from a single institution, which may limit generalizability to other educational contexts with different student populations, teaching methodologies, and institutional cultures. Second, the study focused exclusively on quantifiable academic and behavioral features, potentially overlooking important psychological factors such as motivation, learning style preferences, self-efficacy, and mental health status. Third, the study did not incorporate external contextual factors such as family circumstances, socioeconomic conditions, employment status, or major life events that may impact academic performance.

Fourth, the binary classification approach simplifies the inherent complexity of academic performance, which is multidimensional in nature. Students may excel in some courses while struggling in others, and overall success depends on many variables beyond the scope of this study. Fifth, the temporal dimension was limited as the study utilized cross-sectional data rather than longitudinal data tracking student progress over extended periods. Finally, the study did not examine potential fairness or bias issues in the predictive models, which could have differential impacts on students from different demographic groups.

5.3 Recommendations for Educational Institutions

Based on these findings, the following recommendations are offered to educational institutions:

- Establish data-driven early warning systems using machine learning models to identify at-risk students proactively
- Focus intervention strategies on improving attendance, as it is the most influential factor in academic success
- Implement programs to increase class participation and student engagement in learning activities
- Develop personalized academic support programs tailored to individual student needs and risk profiles
- Monitor and regularly update predictive models to maintain accuracy and adapt to changing student populations

The Asian Thinker
A Quarterly Bilingual Peer-Reviewed Journal for Social Sciences and Humanities
Year-8 Volume: II, April-June, 2026 Issue-30 ISSN: 2582-1296 (Online)
Website: www.theasianthinker.com **Email**: asianthinkerjournal@gmail.com

- Ensure ethical implementation of predictive systems with appropriate safeguards against algorithmic bias

5.4 Future Research Directions

Future research should address the limitations identified in this study through several extensions and improvements:

- Conduct multi-institutional studies involving diverse institutions to develop models with broader applicability
- Incorporate psychological and behavioral measures through validated surveys and assessments
- Develop regression models for continuous GPA prediction to capture nuanced performance variations
- Implement deep learning approaches such as neural networks for capturing complex non-linear patterns
- Conduct longitudinal studies tracking students over multiple semesters and academic years
- Develop real-time early warning systems using streaming data and automated alerts
- Examine fairness and bias in predictive models across different demographic groups

In conclusion, machine learning represents a powerful and promising approach for predicting student academic performance and enabling early intervention. By combining robust algorithms, comprehensive data, and thoughtful implementation, educational institutions can leverage these tools to improve student outcomes and institutional effectiveness.

6. References

1. Romero, C., Espejo, P. G., Zafra, A., Marbán, O., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use an educational web-based system. In Proceedings of educational data mining (pp. 165–172).
2. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In Proceedings of 5th FUture BUsiness TEChnology conference (pp. 5–12).
3. Kabakchieva, D. (2013). Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research, 2(1), 14–19.
4. Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37–66.
5. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

6. Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).
8. Quinlan, J. R. (1993). C4.5: Programs for machine learning. Morgan Kaufmann.
9. Baker, R. S., & Siemens, G. (2014). Educational data mining and learning analytics. In Learning Analytics (pp. 252–278). Springer, New York, NY.
10. Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student success. In ELMAR (Electronic Engineering and Computing), 2012 Proceedings of the 34th International Convention (pp. 426–431). IEEE.
11. Wakelam, E., Coop, R., & Johnson, T. (2018). Predictive analytics for student retention. Journal of Educational Data Mining, 10(2), 45–62.
12. Márquez-Vera, C., Morales, C. R., & Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. Journal of Educational Data Mining, 5(2), 1–7.