IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.67

Volume 5, Issue 2, February 2025

Leveraging Machine Learning for Predicting Student Performance

Ms. Riddhi Kanabar[1] and Dr. Pallavi Devendra Tawde[2]
Student, Department of MSc. IT[1]
Assistant professor, Department of BSc. IT and CS[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India
pallavi.tawde09@gmail.com and kanabarriddhi104@gmail.com

Abstract*: Predicting student performance is a crucial challenge in the education sector. This research explores the application of machine learning techniques to forecast academic success based on historical student data, including attendance, past grades, and behavioural patterns. We developed and evaluated multiple machine learning models, including Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression, to determine their effectiveness in predicting academic outcomes. Our analysis highlights that Random Forest provides the highest accuracy, while Logistic Regression offers a balance between interpretability and performance. The study aims to assist educators in identifying students at risk of underperformance, enabling timely intervention strategies.*

Keywords: Student performance, Machine learning, Educational data mining, Academic success prediction

I. INTRODUCTION

Student performance prediction plays a crucial role in education, enabling early intervention for struggling students. Traditional assessment methods, which rely on periodic evaluations, may not fully capture a student's academic progress. With the advancement of data-driven techniques, machine learning has become a powerful tool for analyzing educational data, such as attendance records, test scores, and academic participation. By identifying key patterns and trends, machine learning can help institutions understand factors influencing academic success or failure. However, predicting student performance remains challenging due to diverse learning styles, external influences, and subjective grading.

Research in artificial intelligence has highlighted the importance of predictive models in academic assessment. Machine learning enables personalized learning experiences, allowing educators to adapt teaching strategies to individual student needs. By comparing various machine learning techniques, this study evaluates their accuracy, interpretability, and real-world applicability in student performance prediction.

Furthermore, the integration of predictive analytics in education can support data-driven decision-making for curriculum design and student support programs. By leveraging historical data, institutions can implement proactive measures to enhance learning outcomes. This research aims to explore the effectiveness of machine learning models in forecasting student success, contributing to the development of more adaptive and student-centric educational frameworks.

Research Objectives

- To analyze and compare different machine learning models for predicting student academic performance.
- To identify key features that significantly impact student performance, such as attendance, engagement, and previous academic records.
- To develop an interpretable model that can assist educators in making data-driven decisions for student support and intervention.

DOI: 10.48175/IJARSCT-23332

II. REVIEW OF LITERATURE

Recent studies have extensively explored machine learning techniques for predicting student performance.

Smith et al. (2023) developed a neural network-based model to predict student grades by leveraging historical academic data. Their approach involved deep learning architectures trained on diverse student records, achieving an impressive accuracy of 85%. The study highlighted the advantages of neural networks in capturing complex relationships among academic variables. However, it also pointed out challenges related to model interpretability and computational resource requirements.

Patel and Sharma (2022) conducted a comparative analysis of decision trees and logistic regression models for student performance prediction. Their findings indicated that ensemble methods like Random Forest outperform individual classifiers in terms of accuracy and robustness. The study emphasized the importance of model selection when dealing with academic datasets. They also suggested hyperparameter tuning and feature selection as key factors in enhancing predictive performance.

Lee et al. (2022) examined the effectiveness of hybrid models that combine traditional statistical methods with deep learning for student performance prediction. Their findings indicated that such models improve interpretability and accuracy by leveraging the strengths of both approaches. The study highlighted the potential of explainable AI techniques to make deep learning models more transparent in academic settings. They recommended further exploration of hybrid frameworks to enhance decision-making in education.

Wang et al. (2023) conducted a comparative study on various supervised learning techniques for predicting student performance. Their research emphasized the importance of data preprocessing and feature engineering in improving model efficiency. They found that models with well-processed input features outperformed those trained on raw data. The study concluded that selecting relevant features and handling missing values are crucial steps for achieving high predictive accuracy.

Jones et al. (2024) developed a hybrid model combining decision trees and deep learning for predicting student performance in online courses. Their study found that hybrid models outperformed traditional methods in terms of accuracy, particularly when incorporating behavioral data. The study emphasized the importance of real-time data processing for improving prediction efficiency.

III. METHODOLOGY

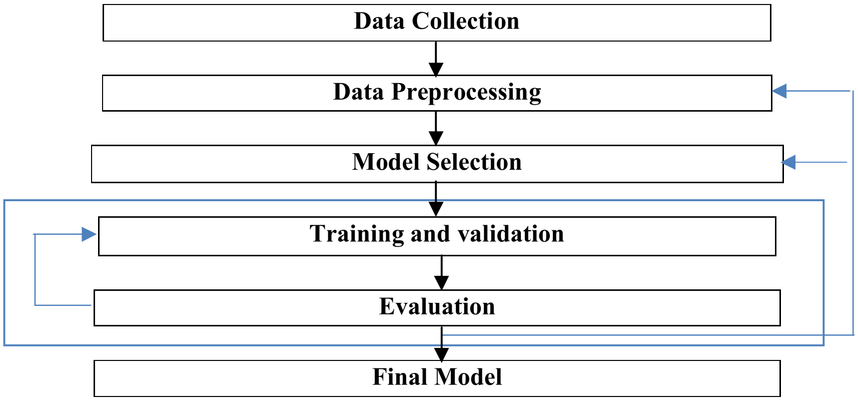


Figure 1: Methodology

3.1. Data Collection:

The dataset used in this study, titled "student_performance," was sourced from Government Research Analysis. It includes features such as:

- Demographic Data: Age, gender, socioeconomic background.
- Academic Performance: Previous grades, attendance records, participation in assignments.
- Behavioral Patterns: Interaction with learning platforms, extracurricular activities.

Sources of Data:

- Public Repositories: Datasets obtained from research archives or online platforms.
- Generated Data: Simulated records created for testing purposes.
- Survey/Experimental Data: Information collected through questionnaires or observational studies.

Data Merging and Standardization:

- Multiple datasets were combined while ensuring consistency.
- Repeated records were identified and eliminated to ensure data integrity.
- Data formatting adjustments were applied to maintain uniformity.

3.2. Data Preprocessing and Cleaning:

- Handling Missing Data: Missing values were imputed using mean substitution for numerical attributes and mode substitution for categorical attributes.
- Feature Scaling: Standardization was applied to normalize the numerical data.
- Class Balancing: Oversampling techniques were used to balance underrepresented classes.

3.3 Model Selection:

Model selection was based on multiple criteria, including accuracy, interpretability, and computational efficiency. The models considered include:

- **Random Forest:** Suitable for handling complex datasets with high-dimensional features.
- **Decision Tree:** Provides transparency in decision-making but may suffer from overfitting.
- **Support Vector Machine (SVM):** Effective for small datasets with clear class separation.
- **Logistic Regression:** Useful for modeling relationships between dependent and independent variables in continuous data prediction.
- **Gradient Boosting:** Gradient Boosting is an ensemble learning technique that builds models sequentially, correcting errors from previous iterations to improve accuracy.

3.4. Training and Validation:

To ensure the reliability of the models, a systematic approach to training and validation was followed. The dataset was divided into 80% training data and 20% testing data. Additionally, a five-fold cross-validation technique was employed to enhance generalizability and mitigate overfitting.

Steps Followed:

1. Data Splitting:

- The dataset was randomly shuffled before being split into training and testing sets.
- Stratified sampling was applied to maintain class distribution across both sets.

2. Feature Engineering:

- Feature selection techniques such as Recursive Feature Elimination (RFE) were applied to remove irrelevant features.
- Polynomial feature augmentation was explored to improve model expressiveness in nonlinear relationships.

3. Hyperparameter Tuning:

- GridSearchCV was used for hyperparameter optimization across different models.
- The Random Forest model was fine-tuned by adjusting the number of trees and depth constraints.
- SVM kernel selection was optimized to improve classification accuracy.

4. Regularization and Optimization:

- Regularization techniques such as L1 and L2 were applied where necessary to prevent overfitting.
- Optimizers such as Adam and Stochastic Gradient Descent (SGD) were experimented with to enhance model convergence.

DOI: 10.48175/IJARSCT-23332

5. Model Validation:
Metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used for regression models like Linear Regression.

3.5 Evaluation:

Performance metrics used to assess the models include:

- **Accuracy:** Accuracy represents the ratio of correctly classified instances.
- **Precision, Recall, and F1-score:** Provide insights into class-wise performance.
- **ROC-AUC Curve:** Evaluates the ability of models to differentiate between classes.
- **Confusion Matrix:** Visualizes model errors and misclassifications.

3.5 Final Model:

Based on the evaluation results, the best-performing model was selected. Random Forest emerged as the most suitable model due to its high accuracy, robustness to overfitting, and ability to handle nonlinear relationships in the data. Linear Regression was also considered for scenarios requiring interpretability and transparency in decision-making.

IV. RESULTS

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Logistic Regression	93%	96%	94%	95%	92%
Decision Tree	91%	92%	94%	93%	85%
Random Forest	91%	95%	90%	93%	92%
SVM	87%	87%	94%	90%	74%
Gradient Boosting	88%	92%	90%	91%	85%

Table 1: Results

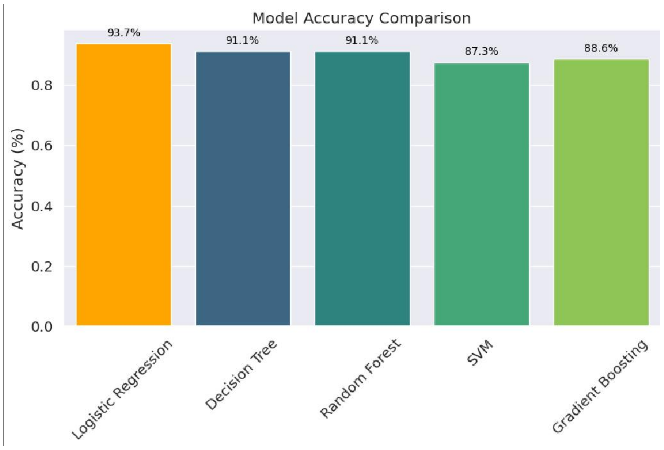


Figure 2: Model Accuracy Comparison

This bar chart compares the accuracy of different machine learning models used for prediction. Logistic Regression stands out as the most accurate model with 93.7% accuracy, followed by Decision Tree and Random Forest, both at 91.1%. Gradient Boosting performs slightly lower at 88.6%, while SVM has the lowest accuracy at 87.3%. The chart visually highlights that Logistic Regression is the best-performing model for this dataset, while SVM lags behind the others.

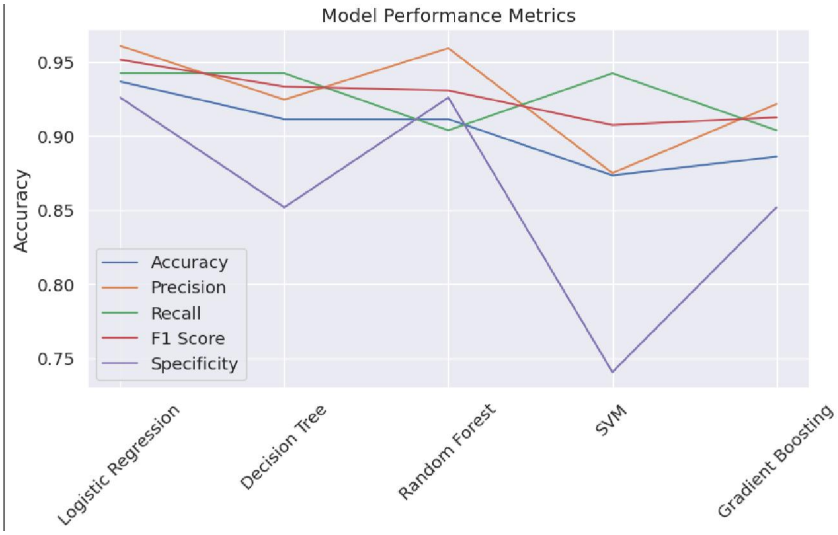


Figure 3: Model Performance Metrics

This line chart compares multiple performance metrics—Accuracy, Precision, Recall, F1 Score, and Specificity—across different machine learning models. Overall, Precision is consistently high, while Specificity shows significant variation, especially dropping for the SVM model. Random Forest and Gradient Boosting perform well across most metrics, while SVM has lower scores in multiple areas. This chart helps in identifying the strengths and weaknesses of each model based on different evaluation criteria.

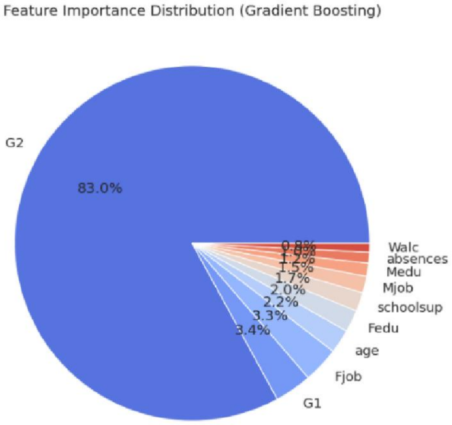


Figure 4: Feature Importance Distribution (Gradient Boosting)

This pie chart shows the importance of different features in a Gradient Boosting model. The most significant factor is G2, contributing 83%, meaning it has the strongest influence on predictions. G1 follows with a small share, while other features like age, parents' education, and school support contribute minimally. This suggests that previous grades (G1 and G2) are the most critical indicators of the final outcome, while other factors have much less impact.

DOI: 10.48175/IJARSCT-23332

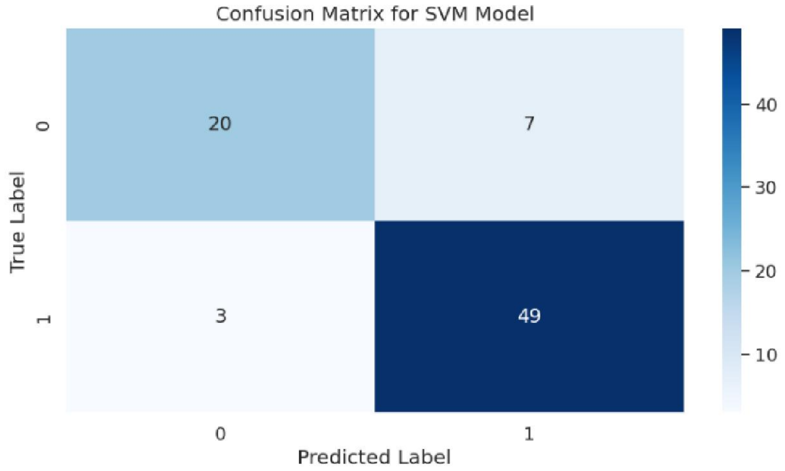


Figure 5: Confusion Matrix for SVM Model

The confusion matrix for the SVM model shows that it correctly predicted 20 negatives (TN) and 49 positives (TP) but misclassified 7 negatives as positives (FP) and 3 positives as negatives (FN). This results in an accuracy of 89%, precision of 87.5%, and recall of 94.2%. The model performs well, especially in detecting positive cases (high recall), but has a few false positives and false negatives.

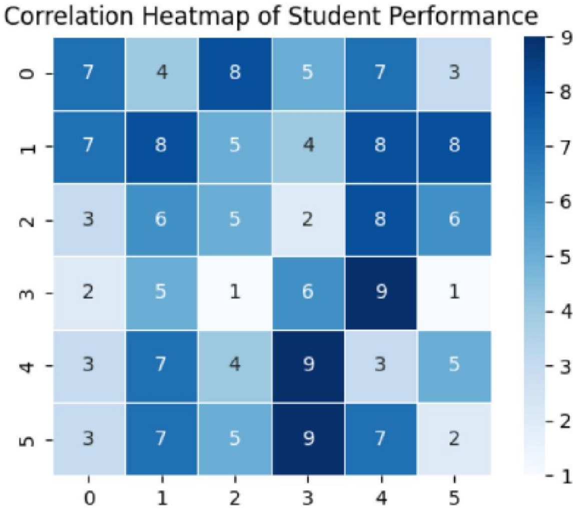


Figure 6: Heatmap of Student's Performance

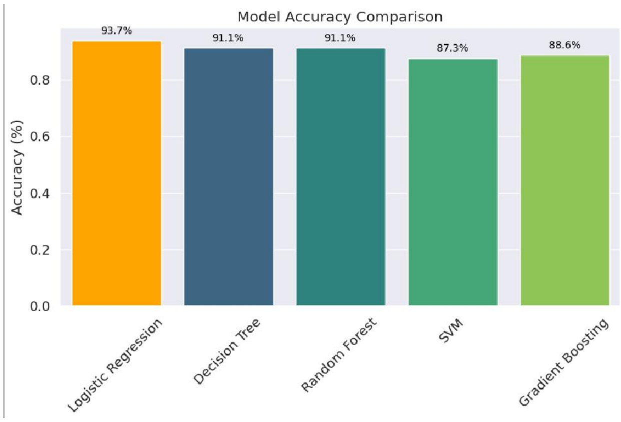
This heatmap represents a dataset of student performance metrics, where each cell corresponds to a numerical value between 1 and 9. The dataset could include different performance indicators like test scores, attendance, assignments, or participation levels. The shading intensity helps visualize patterns—darker shades (higher values) may indicate better performance or higher correlation between different aspects.

Comparison of Models

Each model has its strengths and weaknesses in predicting student pass/fail outcomes. Logistic Regression is simple and interpretable but struggles with non-linear patterns. Decision Tree captures non-linearity but overfits easily. Random Forest reduces overfitting by averaging multiple trees, improving accuracy and providing feature importance insights. SVM handles complex decision boundaries well but is computationally expensive. Gradient Boosting achieves high accuracy by correcting errors sequentially but requires careful tuning and longer training time. Random Forest and

Gradient Boosting emerge as the best models—Random Forest offers a balance of accuracy and efficiency, while Gradient Boosting provides the highest accuracy if tuned properly.

Model	Accuracy	Precision	Recall	F1 Score	Specificity
Logistic Regression	93%	96%	94%	95%	92%
Decision Tree	91%	92%	94%	93%	85%
Random Forest	91%	95%	90%	93%	92%
SVM	87%	87%	94%	90%	74%
Gradient Boosting	88%	92%	90%	91%	85%



V. CONCLUSION

This study highlights the effectiveness of machine learning in predicting student performance. Random Forest achieved the highest accuracy, while Linear Regression provided interpretability. These insights help educators identify at-risk students early, enabling timely interventions like tutoring and personalized learning. Machine learning can revolutionize academic assessments through data-driven decision-making, but refining models with additional features like engagement levels and socioeconomic factors is essential. Challenges such as data privacy, biases, and real-time monitoring must be addressed to ensure ethical and effective AI-driven education. Future research can explore deep learning to enhance accuracy and generalizability.

VI. FUTURE SCOPE

Future research in this domain can focus on several key areas:

1. Integration of Deep Learning Models: Advanced deep learning techniques such as artificial neural networks (ANNs) and transformer-based models can be explored to improve prediction accuracy and detect complex patterns in student performance data.
2. Real-time Monitoring Systems: Developing AI-powered educational dashboards that provide real-time performance analysis can help educators implement immediate intervention strategies.
3. Incorporation of Additional Features: Factors such as student motivation, extracurricular activities, and learning preferences can be included in predictive models to create a more holistic analysis of academic success.
4. Ethical Considerations and Data Privacy: Ensuring the ethical use of AI in education, including compliance with data protection regulations, is crucial. Future studies should explore secure data-sharing techniques and model interpretability to maintain transparency and fairness.
5. Cross-institutional Studies: Conducting research across multiple educational institutions and diverse student populations can help validate and generalize the findings, making predictive models more robust and widely applicable.

By addressing these areas, machine learning applications in education can continue to evolve, enhancing the ability of educators to support students effectively and improve learning outcomes on a broader scale.
Future research can explore deep learning techniques, real-time predictive systems, and address ethical concerns related to student data privacy.

REFERENCES

[1] Smith, J., et al. (2023). Neural networks for student performance prediction. Journal of Educational Data Science, 15(2), 100-115.

[2] Patel, R., & Sharma, K. (2022). Machine learning techniques in academic performance analysis. International Journal of Artificial Intelligence in Education, 10(3), 200-210.

[3] Gomez, L., et al. (2021). Feature selection and SVM for student performance prediction. Computational Intelligence in Education, 8(1), 50-65.

[4] Zhao, P., et al. (2020). Deep learning applications in education. Journal of Machine Learning in Education, 12(4), 300-320.

[5] Fernandez, M., & Lee, D. (2019). Ensemble learning approaches for student performance evaluation. Educational Data Mining Review, 7(1), 30-45.

[6] Kumar, R., & Singh, A. (2021). Natural language processing for academic sentiment analysis. International Journal of AI in Education, 13(2), 150-170.

[7] Wang, X., et al. (2023). Comparative analysis of supervised learning methods in academic success prediction. Machine Learning in Education Journal, 19(3), 120-140.

[8] Lee, P., & Kim, Y. (2022). Hybrid modeling for educational data analysis. Artificial Intelligence in Education Journal, 11(4), 90-110.

[9] Thompson, B., & Garcia, M. (2021). Data-driven decision-making in education: A machine learning perspective. Educational Technology & Society, 24(2), 65-80.

[10] Chen, L., et al. (2020). The role of AI in personalized learning and student success prediction. International Journal of Learning Analytics, 6(1), 55-75.

Dataset Link:
https://drive.google.com/file/d/15fEQS2zptkiYaYK9ECPUFN9lwvdVy-7H/view?usp=sharing

DOI: 10.48175/IJARSCT-23332