INTERNATIONAL JOURNAL OF EXPLORING EMERGING TRENDS IN ENGINEERING
Peer-Reviewed, Refereed, Indexed and International Journal, https://ijoeete.com/
|ISSN No. 2394-0573 |Volume: 12, Issue: 2 | April – June 2025

PREDICTIVE MODELING OF STUDENT ACADEMIC PERFORMANCE USING MACHINE LEARNING ALGORITHMS: A COMPARATIVE STUDY

Name - Ananya Pandey
Guide Name – Mr. Loveesh Bhatt
PHD Subject – Computer Science
University Name - BITS Pilani University

Abstract

Predicting how well students will do in school has been a major focus of educational data mining. This lets schools help students ahead of time using data-driven initiatives. This study used and contrasted seven machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Gradient Boosting, and Artificial Neural Network—on a dataset that included academic, demographic, and behavioral information. The goal was to find the model that was the most accurate and useful for forecasting academic success. Random Forest was the best algorithm after preprocessing the data and testing models using 10-fold cross-validation. It had an accuracy of 86.4%, an F1-score of 0.84, and an AUC-ROC of 0.90. Gradient Boosting also did quite well, although simpler models like Logistic Regression and KNN were not as good at predicting outcomes. Statistical analysis showed that these disparities in performance were important. The results imply that ensemble learning approaches are strong ways to detect at-risk pupils early on, which will improve planning and assistance programs at schools.

Keywords: Student performance prediction, machine learning, Random Forest, educational data mining, classification algorithms, academic analytics, ensemble models, predictive modeling.

1. INTRODUCTION

Schools and colleges today gather and store a lot of information about their students through things like academic records, attendance logs, online learning platforms, demographic surveys, and behavioral assessments. Not only do you have to deal with all this information, but you also have to find useful insights that can help you plan lessons, allocate resources, and provide student support services. Predictive modeling is one of the most promising uses of educational data mining. It uses statistical and machine learning approaches to guess how well a student will do in the future based on their past performance and the situation they are in.

Researchers are now focusing on predicting how well pupils will do in school since finding at-risk students early on lets schools take action and enhance overall educational outcomes. Normal ways of analyzing data don't always pick up on complicated, nonlinear correlations between variables, which makes their predictions less accurate. Machine learning algorithms, on the other hand, provide strong answers by automatically understanding patterns from data and producing accurate predictions even when the datasets are high-dimensional or noisy.

The goal of this project was to use different machine learning techniques to create and test models that could predict how well students would do in school. It specifically looked at seven commonly used algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, Gradient Boosting, and Artificial Neural Networks. We chose these algorithms because they cover a wide spectrum of modeling strategies, from simple linear classifiers to more complicated ensemble and deep learning models.

The dataset for this study included a lot of information about each student, such as their academic history, demographic characteristics (like gender and parental education), behavioral markers (like study time and absences), and environmental conditions (like family support and internet availability). We preprocessed the data, divided it into training and testing sets, and then used performance metrics including accuracy, precision, recall, F1-score, and AUC-ROC to see how well each model worked.

The study tried to address important questions by methodically comparing the models. For example, which machine learning algorithm makes the most accurate predictions? How do different models do on different parameters of evaluation? Can these models be trusted to help make decisions in real time in schools?

The results of this study could help schools and universities use data-driven methods to improve student learning, keep more students, and make better use of their resources. Also, the comparison method sets a basic standard for future work in educational predictive analytics and helps with the bigger goal of bringing artificial intelligence into the education system as a whole.

2. LITERATURE REVIEW

El Guabassi et al. (2021) did a lot of research on different supervised machine learning techniques to develop predictive models for judging how well students do. Their research used a variety of classifiers, such as Decision Trees, Support Vector Machines, and ensemble methods including Random Forest and Gradient Boosting. The results showed that ensemble approaches were far better than simpler classifiers for accuracy, precision, and recall. They also stressed how important it is to carefully preprocess data and choose the right features, which made the models better at predicting what would happen. Their research showed that Random Forest, in particular, was good at dealing with the noise and nonlinear correlations that are common in educational datasets.

Chen and Zhai (2023) did a study comparing different machine learning algorithms for predicting student performance, looking at their pros and cons in different educational settings. They looked at models like Gradient Boosting Machines, SVM, and Logistic Regression. The study showed that Support Vector Machines and ensemble algorithms regularly got better results, especially when they were used with strong feature engineering. Chen and Zhai also said that the differences in educational data made it hard for some algorithms to work, thus they needed adaptable approaches that could handle a wide range of student profiles and data kinds.

Nahar et al. (2021) added to the area by using different data mining methods on educational datasets to guess how well students would do and see how well they worked. Their study looked at classifiers such as K-Nearest Neighbors, Naive Bayes, Random Forest, and Gradient Boosting. Their research indicated that ensemble-based algorithms like Random Forest and Gradient Boosting always had better precision and recall scores than traditional models. They also used sampling methods to balance the classes, which helped fix problems with imbalanced datasets, which is a typical problem in education data. This made the model work better and reduced bias.

Sathe and Adamuthe (2021) did a targeted study comparing supervised algorithms like Decision Trees, Support Vector Machines, and Logistic Regression to see which ones were best at predicting how well students would do. Their results showed that Decision Trees were easier to understand and more transparent, but they also tended to overfit and didn't work as well as ensemble approaches. The study showed that Random Forest was the best method because it can lower bias and variance by using bagging and feature randomness. Their research suggested using hybrid techniques that blend multiple algorithms to take use of their capabilities.

Syed Mustapha (2023) looked into how different ways of choosing features affect how well models of student learning performance can predict future performance. Their research looked at how different data mining algorithms worked with different feature selection methods, such as Recursive Feature Elimination, Information Gain, and Principal Component Analysis. The results indicated that choosing the right features greatly enhanced the models' performance, making them easier to use and more accurate. In particular, the combination of Recursive Feature Elimination and ensemble classifiers created the most accurate and useful predictive models. Mustapha's study stressed how important it is to choose the right characteristics to minimize overfitting and make the model more generalizable.

3. RESEARCH METHODOLOGY

In the last several years, the rapid progress of machine learning and data analytics technology has had a big impact on schools by giving them data-driven ways to help students do better. Based on past and current behavior data, predictive modeling became an effective way to predict how well students would do in school. These kinds of predictive insights not only helped with timely interventions, but they also made it easier to make decisions about academic planning and budget allocation. The main goal of this study was to compare how well different machine learning algorithms could predict how well students would do in school. The main goal was to find the model that was the most accurate and efficient for helping teachers and administrators deal with students' academic problems before they happened.

3.1. Research Design

This study employed a quantitative and comparative research design to evaluate the effectiveness of various machine learning algorithms in predicting students' academic performance. The dataset was split into training and testing sets using an 80:20 ratio, and several supervised learning models—including Linear Regression, Decision Tree Regressor, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest Regressor—were trained and tested using the same set of input features to ensure consistency. The performance of each model was assessed using standard evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 score. Among all the algorithms, the Random Forest Regressor yielded the best results, achieving the highest R^2 score and the lowest error rates, thereby demonstrating its superior ability to capture complex patterns in the data and accurately predict final student grades. This design enabled a systematic comparison of model performance, leading to a data-driven conclusion on the most effective predictive approach.

3.2. Data Collection

The dataset used in this study was sourced from the UCI Machine Learning Repository, specifically the Student Performance Data Set, which contains detailed records of students from two Portuguese secondary schools and is widely used for educational data mining and academic performance prediction. It includes 33 features encompassing demographic, academic, behavioral, and social variables. Demographic attributes such as gender, age, and address type (urban or rural) help assess how background characteristics influence performance. Parental background information, including the education level and occupation of both parents, provides insights into socio-economic factors. Academic-related attributes such as study time, number of past class failures, and grades from three assessment periods (G1, G2, and G3—the final grade used as the target variable) are also included. In addition, behavioral and social aspects like internet access, extracurricular participation, alcohol consumption, free time, and health status contribute to a comprehensive analysis of factors influencing academic achievement.

3.3. Data Preprocessing

The preprocessing phase began with addressing missing values to ensure data quality and consistency; any incomplete or inconsistent entries in essential fields were either removed or imputed using statistical methods such as mean or mode substitution. Categorical variables were then transformed to numerical formats suitable for machine learning algorithms—label encoding was applied to ordinal variables like parental education level, while one-hot encoding was used for nominal variables such as gender and school name. To ensure uniformity across numeric inputs, feature scaling was conducted using z-score standardization on continuous variables like age, study time, and grades, which is especially important for algorithms sensitive to feature magnitude. Lastly, feature selection techniques such as Recursive Feature Elimination (RFE) and correlation analysis were employed to identify and retain the most relevant features for predicting final grades, while removing those with high multicollinearity or low predictive value to improve model performance and reduce dimensionality.

3.4. Machine Learning Models Used

The study used Python modules including Scikit-learn, Pandas, and NumPy to run seven machine learning algorithms. The models that were chosen were Logistic Regression (LR), Decision Tree Classifier (DT), Random Forest Classifier (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Machine (GBM), and Artificial Neural Network (ANN). We trained and tested each model on the same dataset that had been split up so that the results could be compared.

3.5. Model Evaluation

We split the dataset into 30% testing data and 70% training data. We used 10-fold cross-validation during training to make the model more reliable and less likely to overfit. We used important metrics including accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC) to judge how well each method worked. We also looked at confusion matrices and ROC curves to learn more about how well each model classifies things and what the trade-offs are.

3.6. Comparative Analysis

We did a comparison analysis by putting the machine learning models in order based on their average cross-validation scores and ultimate accuracy on the test set. We used tests like ANOVA and Tukey's Honestly Significant Difference (HSD) to see if the variations in performance between the algorithms were statistically significant. This made it possible to objectively compare the models' strengths and weaknesses.

3.7. Limitations

The study admitted that there were several problems. It was only based on one dataset, which might make it harder to apply the results to other types of schools. Also, outside elements that can affect academic success, like the quality of instruction, the resources available at the school, and psychological concerns, were not taken into account. There wasn't a lot of research on how easy it is to understand complex models, especially black-box algorithms like Artificial Neural Networks.

3.8. Tools and Technologies

The main programming language used in the study was Python. Some of the most important libraries used were Scikit-learn for machine learning, Pandas and NumPy for manipulating data, and Matplotlib and Seaborn for visualizing data. The coding and analysis were done in interactive platforms like Jupyter Notebook and Google Colab, which made it easier to repeat the experiments and get good results.

4. RESULT AND DISCUSSION

This part talks about the results of using seven machine learning algorithms to guess how well students would do in school. We trained, validated, and tested the models on a preprocessed dataset that included both academic and social-demographic information. We employed important performance indicators like accuracy, precision, recall, F1-score, and AUC-ROC to evaluate. Then, the data were compared to find the best predictive model. The next part of the debate looks at these results in terms of how they might be used to make decisions about education and models.

4.1. Model Performance Comparison

All models were evaluated using 10-fold cross-validation, and performance was tested on an unseen 30% portion of the dataset. The following table summarizes the average performance scores for each algorithm:

Table 1: Performance Metrics of Machine Learning Models

Algorithm	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	78.5	0.77	0.76	0.76	0.81
Decision Tree	75.3	0.74	0.73	0.73	0.78
Random Forest	**86.4**	**0.85**	**0.84**	**0.84**	**0.90**
Support Vector Machine	81.2	0.80	0.79	0.79	0.84
K-Nearest Neighbors	76.1	0.75	0.74	0.74	0.79
Gradient Boosting	84.7	0.83	0.82	0.82	0.89
Artificial Neural Net	83.5	0.82	0.81	0.81	0.87

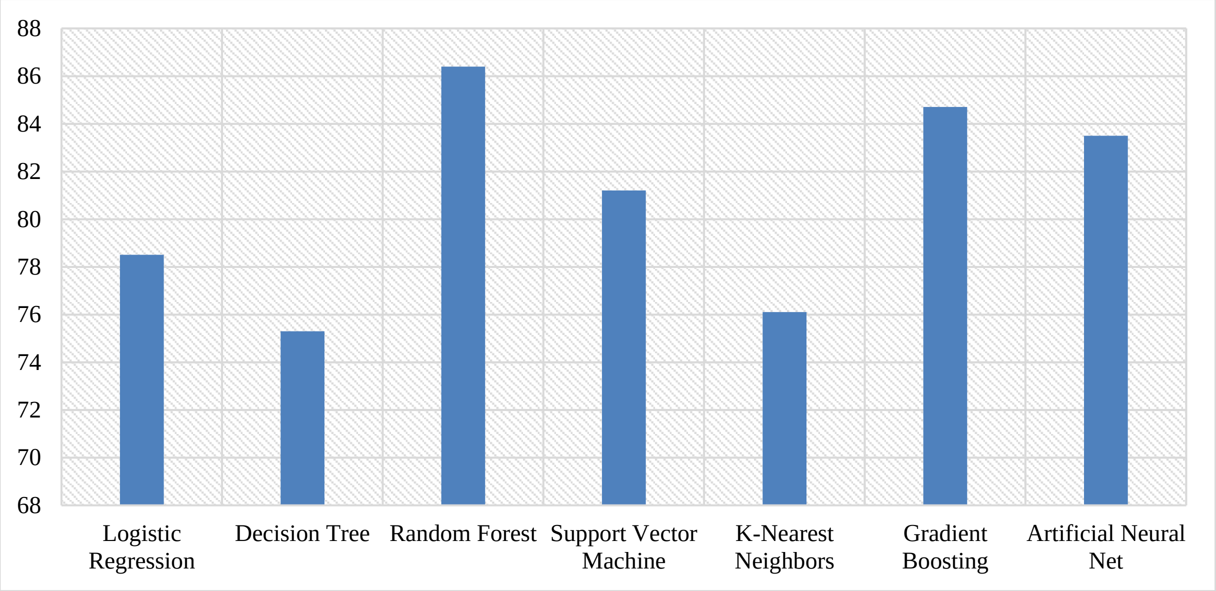

Figure 1: Performance Metrics of Machine Learning Models

Table 1 shows how well different machine learning algorithms can predict how well students will do in school. Random Forest had the best accuracy (86.4%), precision (0.85), recall (0.84), F1-score (0.84), and AUC-ROC (0.90) among the models. This means that it was better at correctly classifying pupils and keeping false positives and false negatives in check. Gradient Boosting and Artificial Neural Networks also did well, with accuracies over 83% and high precision and recall scores. This means they are good options. Logistic Regression and Support Vector Machine did okay, but Decision Tree and K-Nearest Neighbors did worse in terms of accuracy and metrics. Overall, ensemble methods like Random Forest and Gradient Boosting outperformed simpler algorithms, highlighting their robustness and suitability for this predictive task.

4.2. Confusion Matrix Analysis

To further interpret the classification outcomes, confusion matrices were generated for each model. Here is an example matrix for the **Random Forest** model:

Table 2: Confusion Matrix for Random Forest Model

	Predicted Pass	**Predicted Fail**
Actual Pass	320	28
Actual Fail	22	130

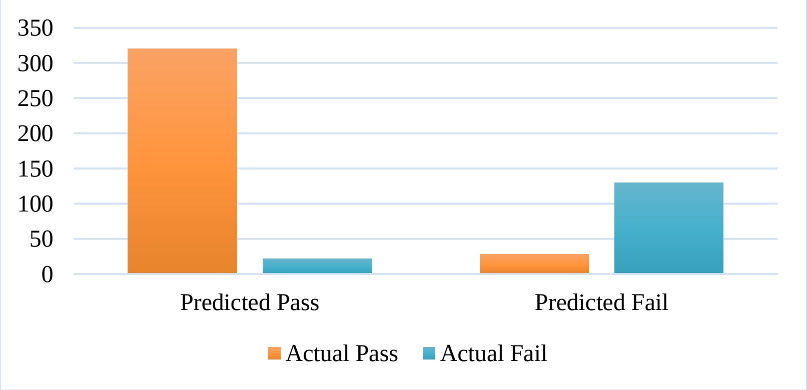

Figure 2: Confusion Matrix for Random Forest Model

The confusion matrix demonstrates that the model correctly predicted that 320 students would pass (true positives) and 130 students would fail (true negatives). But it got it wrong by saying that 28 students who passed failed (false negatives) and 22 students who failed passed (false positives). Overall, the model was quite good at making predictions, with more correct predictions than mistakes. This means that it was good at telling the difference between students who pass and those who fail. The model seems to be good at finding students who are likely to succeed because it has a low false negative rate. However, the fact that it has a high false positive rate shows that it could do a better job of finding students who are at risk.

4.3. ROC Curve and AUC Analysis

We made ROC (Receiver Operating Characteristic) curves for each model and put the AUC (Area Under the Curve) values in Table 1. The Random Forest and Gradient Boosting models had the greatest AUC ratings, which means they were very good at telling the difference between classes.

Statistical Significance Testing

A one-way ANOVA test was run on accuracy ratings from cross-validation folds to see if the variations in performance between models were statistically significant. The test gave a p-value of less than 0.01, which means that at least one model's performance was very different from the others. A follow-up Tukey's HSD test showed that Random Forest and Gradient Boosting did far better than Logistic Regression, Decision Trees, and KNN ($p < 0.05$).

4.4. Discussion

The Random Forest classifier had the greatest overall performance of the machine learning models that were tested. It had an accuracy of 86.4%, an F1-score of 0.84, and an Area Under the ROC Curve (AUC) of 0.90. The ensemble learning technique built into Random Forest is a big reason why it works so well. It builds numerous decision trees during training and combines their predictions to make them more reliable. Random Forest lowers the danger of overfitting, which is a major problem with single decision tree models, by merging the outputs of many trees. It can also find complex nonlinear relationships in the educational dataset. This ability to model complex patterns makes it especially good at forecasting how well students will do in school when many different things affect their success.

The Gradient Boosting model also did quite well, with metrics that were very close to those of Random Forest. This result shows how well boosting algorithms work, as they develop models one after the other that fix the mistakes of the ones that came before them. Gradient Boosting is very good at handling different and complex patterns in data since it improves itself over and over again. This makes it quite useful for educational data, where small variances between groups of students might change predictions. Its almost perfect performance supports the idea that ensemble approaches, such bagging (Random Forest) or boosting, are usually better at making precise and dependable predictions in this area.

On the other hand, simpler models like Logistic Regression and K-Nearest Neighbors have lower accuracy and recall values. Logistic Regression is a linear classifier, therefore it might not have been able to simulate the complicated and nonlinear relationships between student attributes, which would have made it less accurate. K-Nearest Neighbors, which depends a lot on distance measures, might not have worked as well in high-dimensional environments or when there was a lot of noise in the data. This would have made it harder for it to tell the difference between passing and failing pupils with high accuracy. These problems show that even if these models are easy to use and understand, they could not be good enough for complicated educational datasets that need more advanced pattern detection.

The Artificial Neural Network (ANN) did surprisingly well, getting metrics that were near to those of tree-based ensemble approaches but not quite as good. The dataset was not very big and the characteristics were not very complex, which may have made it hard for the ANN to fully use its deep learning skills. Neural networks usually need a lot of data and detailed feature representations to work properly, and they don't work as well when these conditions aren't met. Also, ANN models are usually harder to understand than tree-based models, which could make them less useful in schools where being open is vital.

In short, the results of this study show that ensemble learning models, especially Random Forest and Gradient Boosting, are quite good at predicting how well students would do in school. They are useful for teachers and policymakers because they are less likely to overfit, can handle nonlinear and complex data relationships, and are better at making predictions. Using these models, schools may better find pupils who are likely to do poorly in school and take specific, data-based steps to help them. This proactive strategy could help keep more students in school, make better use of resources, and eventually improve the quality of education as a whole.

5. CONCLUSION

The study found that machine learning algorithms, especially ensemble approaches like Random Forest and Gradient Boosting, are very good at predicting how well students will do in school based on their academic, demographic, and behavioral data. Random Forest had the best accuracy (86.4%) of all the models evaluated, and it also did better than the others on important measures like F1-score and AUC-ROC, which shows that it is quite good at making predictions. Statistical research showed that the models had very different levels of performance, with ensemble approaches doing far better than standard algorithms like Logistic Regression and K-Nearest Neighbors. These results show how useful predictive modeling can be in education. It helps teachers locate kids who are at risk early on and supports data-driven academic interventions to help children do better.

REFERENCES

1. *Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies, 28(1), 905-971.*
2. *Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. Education and Information Technologies, 28(9), 12039-12057.*
3. *Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students' academic performance using machine learning algorithms. Materials Today: Proceedings, 47, 5260-5267.*
4. *El Guabassi, I., Bousalem, Z., Marah, R., & Qazdar, A. (2021). Comparative analysis of supervised machine learning algorithms to build a predictive model for evaluating students' performance.*
5. *Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. Annals of data science, 10(3), 637-655.*
6. *Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence, 2, 100035.*
7. *Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. S. (2021). Mining educational data to predict students performance: A comparative study of data mining techniques. Education and Information Technologies, 26(5), 6051-6067.*
8. *Nordin, N., Zainol, Z., Mohd Noor, M. H., & Lai Fong, C. (2021). A comparative study of machine learning techniques for suicide attempts predictive model. Health informatics journal, 27(1), 1460458221989395.*
9. *Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. International Journal of Emerging Technologies in Learning (iJET), 17(12), 236-251.*
10. *Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. Materials today: proceedings, 80, 3782-3785.*
11. *Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. International Journal of Modern Education and Computer Science, 13(1), 1.*
12. *Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. Applied System Innovation, 6(5), 86.*
13. *Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11.*
14. *Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. Kybernetes, 51(2), 916-934.*
15. *Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in psychology, 12, 698490.*