Comparative Study of Machine Learning Models for Student Performance prediction

[1] Rohan Sharma, [2] Jabulani Khumalo
[1] Indian Institute of Technology (IIT) Bombay, Mumbai, India
[2] University of South Africa, Pretoria, South Africa
Corresponding E-mail: rohan.sharma@interviauniversity.com

Abstract

Predicting how well students will perform is becoming an increasingly useful application of machine learning in education. Schools, colleges, and universities collect different types of student information, such as attendance, previous grades, study habits, participation, and academic activities. When this information is analyzed properly, it can help teachers and institutions recognize students who may be struggling and provide support before their performance drops further. Traditional approaches to evaluating students usually rely on examinations, teacher observations, assignments, attendance records, and previous academic results. These methods are valuable, but they may not always capture the relationships between several factors at the same time. For example, a student may have reasonable examination results but still be at risk because of declining attendance or reduced participation. Machine learning can help by finding patterns in historical student data and using those patterns to estimate future academic performance. This research presents a comparative study of four machine learning models: Linear Regression, Decision Tree, Random Forest, and Support Vector Regression (SVR). The proposed methodology includes data collection, data preprocessing, exploratory analysis, feature selection, model training, and performance evaluation. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) are used as evaluation measures. The study also discusses the strengths, limitations, and practical uses of each model. Factors such as previous grades, attendance, study time, and assignment performance are considered particularly relevant. Overall, machine learning can provide useful information about student performance and can support educational institutions in taking early and more targeted action when students appear to be at risk.

Keywords: Student Performance, Machine Learning, Educational Data Mining, Linear Regression, Decision Tree, Random Forest, Support Vector Regression

1. Introduction

Education has a major influence on the personal, social, and economic development of individuals and communities. For educational institutions, improving student achievement is not only about delivering course content; it also involves understanding students' needs and recognizing difficulties early. One of the challenges faced by teachers and administrators is identifying students who may struggle academically before the problem becomes serious[1].

Student performance is affected by many different factors. Previous grades, attendance, study time, assignment completion, classroom participation, access to learning resources, family background, motivation, and interaction with teachers can all play a role. These factors do not always work independently[2]. In many cases, a combination of several factors may explain why a student performs well or poorly. This makes it difficult to rely on a single indicator when trying to predict future academic outcomes[3].

Traditionally, student performance is assessed through examinations, assignments, attendance records, and teacher observations. These methods provide important information and should continue to play a central role in education. However, they may not reveal less obvious patterns in large amounts of student data. For instance, a student whose grades are currently acceptable may still be moving toward poor performance if attendance and participation have been decreasing over time [4].

Machine learning offers a way to analyze educational data in a more systematic manner. Instead of looking at each factor separately, machine learning models can learn patterns from historical records and use them to predict academic outcomes for new students. Such predictions can help teachers identify students who may benefit from tutoring, additional learning materials, feedback, or counseling [5].

Student performance prediction can be approached as either a classification or regression problem. Classification assigns students to categories such as high, medium, or low performance. Regression, on the other hand, predicts a numerical value such as a final examination score or overall grade. This study focuses on the regression approach because it provides a continuous estimate of academic performance[6].

Four models are considered in this research. Linear Regression is used as a simple and interpretable baseline. Decision Tree regression can represent nonlinear relationships through a series of understandable rules. Random Forest combines many decision trees to produce more stable predictions and handle complex patterns. Support Vector Regression can model nonlinear relationships through suitable kernel functions, although it generally requires careful parameter tuning and feature scaling [7].

The main objective of this research is to compare these four models for student performance prediction. The study focuses on the major stages of developing a prediction system, including data preprocessing, feature selection, model development, evaluation, and interpretation. It also considers the practical and ethical issues that arise when machine learning is used in education[8].

2. Methodology

This study follows a supervised machine learning regression approach to predict students' academic performance using available academic, behavioral, and demographic information. The research process includes data collection, data preprocessing, exploratory data analysis, feature selection, dataset splitting, model training, prediction, model evaluation, and comparative analysis. The study uses four machine learning models: Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression (SVR)[9]. The dataset may contain features such as previous examination scores, attendance, study time, assignment performance, class participation, and access to learning resources. During preprocessing, missing values, duplicate records, inconsistent entries, and categorical variables are handled appropriately, while numerical features may be scaled where required, particularly for SVR. Feature selection is performed to identify the variables that contribute most to predicting academic performance [10]. The dataset is divided into training and testing sets so that the models can be evaluated on previously unseen student records. Linear Regression provides a simple and interpretable approach, while Decision Trees can represent decision-based relationships. Random Forest combines multiple trees to capture complex and nonlinear patterns, whereas SVR can model more complicated relationships using suitable kernels. Finally, the performance of all four models is evaluated and compared using appropriate regression metrics to identify the most effective model for predicting student academic performance [11].

3. Results and Discussion

The comparative analysis shows that each of the four models offers a different way of approaching student performance prediction. Since the actual performance of a model depends on the dataset and experimental settings, no single algorithm should automatically be considered the best for every educational environment [12].

Linear Regression is a useful baseline because it is simple and relatively easy to interpret. It can reveal broad relationships between variables such as previous grades, attendance, and predicted academic scores. However, its linear structure limits its ability to represent complicated interactions.

Decision Tree regression is more flexible because it can model nonlinear relationships and create clear decision rules. It may reveal combinations of factors associated with different levels of performance. At the same time, an unrestricted tree can become overly complex and overfit the training data[13].

Random Forest reduces some of the weaknesses of an individual Decision Tree by combining multiple trees. As a result, it can provide more stable predictions and handle complex relationships between student characteristics. Its feature-importance information can also be useful when trying to understand which factors matter most [14].

Support Vector Regression provides another flexible option, particularly when nonlinear relationships are present. Its performance can be strong when the model parameters and feature scaling are chosen appropriately. However, it can require more tuning than the simpler models.

A summary of the main characteristics of the models is presented below:

Model | Main Strength | Main Limitation

Linear Regression | Simple and interpretable | Limited ability to model nonlinear relationships

Decision Tree | Easy to understand and flexible | Can overfit the training data

Random Forest | Robust and effective for complex patterns | More complex and less interpretable

Support Vector Regression | Can model nonlinear relationships | Requires careful scaling and parameter tuning

3.1 Evaluation Metrics

Several evaluation metrics can be used to compare the prediction models.

Mean Absolute Error (MAE) represents the average absolute difference between predicted and actual student scores. A lower MAE means that the model's predictions are, on average, closer to the actual values.

Mean Squared Error (MSE) calculates the average of the squared prediction errors. Because the errors are squared, larger mistakes have a greater effect on this metric.

Root Mean Squared Error (RMSE) is the square root of MSE. It is particularly useful because it is expressed in the same units as the predicted score, which makes it easier to interpret.

R-squared (R^2) indicates how much of the variation in the target variable can be explained by the model. In general, a higher R^2 suggests that the model explains more of the variation in student performance, although it should always be considered alongside other evaluation measures.

Using several metrics instead of only one gives a more balanced view of model performance. For example, two models may have similar MAE values but differ in how they handle larger prediction errors.

3.2 Important Factors Affecting Student Performance

Feature importance is an important part of student performance prediction. Previous academic performance is often expected to be one of the strongest predictors because it provides information about a student's existing knowledge and learning progress.

Attendance can also be relevant. Students who attend classes regularly have more opportunities to receive instruction, ask questions, participate in discussions, and receive feedback. Nevertheless, attendance alone does not guarantee strong academic performance.

Study time may also be related to academic outcomes. Students who spend more time studying may have more opportunities to practice and review course material. However, the quality and effectiveness of study are important, meaning that the number of hours alone should not be treated as a complete measure of learning.

Assignment and coursework performance can provide another useful source of information. Unlike a single final examination, assignments can provide repeated evidence of a student's progress throughout the semester. This can make them useful for identifying changes in performance earlier [15].

3.3 Educational Use and Early Intervention

One of the most practical benefits of student performance prediction is the possibility of early intervention. If a model suggests that a student may be at risk of poor performance, educators can respond by offering tutoring, additional learning resources, feedback, mentoring, or academic counseling.

However, machine learning predictions should not be treated as permanent labels. A prediction is only an estimate based on the information available to the model at a particular time. Students can improve or decline depending on changes in their circumstances, motivation, learning strategies, and support systems.

For this reason, prediction systems should be used to support educational decisions rather than to make decisions automatically. Teachers can combine model outputs with their own knowledge of the student and the classroom situation.

3.4 Privacy and Ethical Considerations

Educational data often contains sensitive information about students. Academic records, demographic characteristics, attendance, and behavioral information should therefore be collected and stored responsibly. Institutions need to apply appropriate privacy and security practices when developing student prediction systems.

Bias is another important concern. If historical data contains unfair patterns related to socioeconomic background, access to resources, or other characteristics, a machine learning model may learn and reproduce those patterns. Model evaluation should therefore include checks for fairness and possible differences in performance across student groups.

Another challenge is that student behavior changes over time. A student who performs poorly during one semester may improve after receiving support. Similarly, a student who performs well may later face difficulties. Prediction systems should therefore be updated regularly and should not rely permanently on outdated information.

Interpretability is particularly valuable in educational applications. Teachers and administrators may want to know why a student has been identified as potentially at risk. Models that provide useful information about important features can make it easier for educators to understand predictions and decide whether an intervention is appropriate.

3.5 Overall Comparison

Overall, the comparison suggests that model selection should depend on the characteristics of the dataset and the purpose of the prediction system. Linear Regression is attractive when simplicity and interpretability are priorities. Decision Trees are useful when understandable decision rules are important. Random Forest is suitable when the goal is to capture complex relationships while maintaining strong predictive performance. SVR can be useful when nonlinear patterns are present and sufficient effort can be devoted to model tuning.

There is no guarantee that the same model will perform best on every student dataset. The size and quality of the data, feature relationships, preprocessing methods, model parameters, and evaluation strategy can all affect the final results.

Machine learning should therefore be viewed as a support tool rather than a replacement for teachers. Academic performance can be influenced by personal, social, emotional, and environmental factors that may not be recorded in a dataset. Human judgment remains essential when interpreting predictions and deciding how to support students[16].

4. Conclusion

This research presented a comparative study of four machine learning models for predicting student performance: Linear Regression, Decision Tree, Random Forest, and Support Vector Regression. The study described the main stages involved in developing a student performance prediction system, including data collection, preprocessing, feature selection, model training, and evaluation. Each model has its own strengths and weaknesses. Linear Regression provides a straightforward and interpretable baseline. Decision Tree regression can capture nonlinear relationships using understandable decision rules but may overfit if it becomes too complex. Random Forest improves robustness by combining multiple trees and can handle complex interactions between features. Support Vector Regression offers a flexible way to model nonlinear relationships but requires careful scaling and parameter tuning. Overall, machine learning has considerable potential in educational data analysis and student performance prediction. When used responsibly, it can help institutions move from simply reacting to poor academic results toward identifying possible difficulties earlier and providing more personalized support. Future research could strengthen these systems by including real-time learning behavior, online learning activities, natural language processing of student responses, and more advanced machine learning methods. Testing the models on larger and more diverse real-world datasets would also help determine how well the findings generalize across different educational settings.

References

[1] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory,* vol. 55, pp. 1-9, 2015.

[2] R. Ramadugu, "A Formalized Approach to Secure and Scalable Smart Contracts in Decentralized Finance," in *2025 International Conference on Engineering, Technology & Management (ICETM)*, 2025: IEEE, pp. 1-6.

[3] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in *2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT)*, 2018: IEEE, pp. 1-4.

[4] A. Mohammed, "Transforming SOC Operations: Harnessing the Power of AI and ML for Enhanced Threat Detection," *INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY Monthly Peer-Reviewed, Refereed, Indexed,* vol. 8, pp. 11-21, 2024, doi: 10.2017/IJRCS/ICCDMP01.

[5] A. Mohammed, "Artificial Intelligence-Driven Cybersecurity Crimes: Emerging Threats, AI Bots, and Advanced Defense Strategies," *Euro Vantage journals of Artificial intelligence,* vol. 3, no. 1, pp. 46-65, 2026, doi: https://doi.org/10.65923/c2sxrx21.
[6] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time credit card fraud detection using machine learning," in *2019 9th international conference on cloud computing, data science & engineering (Confluence)*, 2019: IEEE, pp. 488-493.
[7] J. Barach, "Enhancing Ransomware Resilience in Cloud-Based HR Systems through Moving Target Defense," *Computers, Materials and Continua,* vol. 86, no. 2, pp. 1-23, 2025, doi: https://doi.org/10.32604/cmc.2025.071705.
[8] S. Malik, S. Harous, and H. El-Sayed, "Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women," in *International Symposium on Modelling and Implementation of Complex Systems*, 2020: Springer, pp. 95-106.
[9] F.-Z. Nakach, A. Idri, and E. Goceri, "A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification," *Artificial Intelligence Review,* vol. 57, no. 12, p. 327, 2024.
[10] A. Qatawneh, "The influence of data mining on accounting information system performance: a mediating role of information technology infrastructure," *Journal of Governance and Regulation/Volume,* vol. 11, no. 1, pp. 141-151, 2022, doi: https://doi.org/10.22495/jgrv11i1art13.
[11] S. Khairnar, "Application of Blockchain Frameworks for Decentralized Identity and Access Management of IoT Devices," *International Journal of Advanced Computer Science & Applications,* vol. 16, no. 6, 2025.
[12] A. M. Qatawneh, "The effect of electronic commerce on the accounting information system of Jordanian banks," *International Business Research,* vol. 5, no. 5, p. 158, 2012, doi: 10.5539/ibr.v5n5p158.
[13] E. Ahmed, "Student performance prediction using machine learning algorithms," *Applied computational intelligence and soft computing,* vol. 2024, no. 1, p. 4067721, 2024.
[14] J. Barach, "Federated Learning for Privacy-Preserving Employee Performance Analytics," *IEEE Access,* vol. 13, 2025, doi: 10.1109/ACCESS.2025.3591360.
[15] D. Bodra and S. Khairnar, "Comparative performance analysis of modern NoSQL data technologies: Redis, Aerospike, and Dragonfly," *arXiv preprint arXiv:2510.08863,* 2025.
[16] M. Beseiso, "Enhancing student success prediction: A comparative analysis of machine learning technique," *TechTrends,* vol. 69, no. 2, pp. 372-384, 2025.