International Journal of Multidisciplinary Research and Development
www.allsubjectjournal.com
ISSN Online: 2349-4182
Received: 10-12-2025, Accepted: 20-01-2026, Published: 07-02-2026
Volume 13, Issue 1, 2026, Page No. 235-238

Machine learning in education: A comparative study on predicting students academic performance

Aparna Shivaji Gaikwad
Rajmata Jijau Shikshan Prasarak Mandal's, Institute of Computer and Management Research, Dudulgaon, Pune, Maharashtra, India

Abstract
This study looks into how machine learning methods can be used to guess how well students will do in school by looking at things like test scores, attendance, and how engaged they are in school. We used and compared four well-known algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) to see which ones worked best. The study shows that Random Forest gave the most reliable outcomes, pointing out that attendance and past scores are important factors that affect success. The results show that predictive analytics can help teachers find students who are likely to do poorly in school and come up with specific ways to help them. This study shows that data-driven methods are becoming more and more useful in education to help students learn better and help schools make better decisions.

Keywords: Artificial intelligence, machine learning, student performance prediction, classification, educational data mining

Introduction

This is possible because of the fast growth of educational data and improvements in artificial intelligence. This has opened up new ways to better understand and improve student learning. However, academic success is affected by many things that are linked to each other, such as past performance, attendance, and participation in class. Even so, academic ability is still one of the best ways to tell if a student will be successful. Traditionally, tests, teacher notes, or regular assessments have been what educators have used. However, these ways often fail to find students who are having trouble early enough so that they can get help right away.

By getting predictive insights from datasets that are both complicated and multidimensional, machine learning opens up new ways to do things. Machine learning is different from standard statistical models because it can see patterns that don't follow a straight line and how variables affect each other. This gives us a more complete picture of how well students are doing. Choice Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) are some of the methods that are being used more and more in educational data mining. Each one has its own benefits: Decision Trees can be read, Random Forests can stop overfitting, Support Vector Machines can handle high-dimensional data well, and KNN provides an instance-based method that is easy to use and very effective.

As part of this study, these algorithms will be used to see how well they can predict how those kids will do in school. We think that by doing this, we will not only be able to find the best methods, but we will also be able to give useful information that can be used to improve teaching methods and ways to help students.

Review of Literature

A lot of research has been done on how to use machine learning to predict how well students will do in school, but the results have been mixed based on the dataset and the situation.

1. Alamri *et al.* (2020) [2] used Random Forest and SVM to guess how well Portuguese students would do in language and math classes. They found that both models were pretty good at what they did. Their work showed how important it is to use more than just numbers and include behavioural data as well.
2. Orji and Vassileva (2022) [4] showed that tree-based models like Random Forest can accurately catch both cognitive and motivational factors, getting very close to 95% of the time. Their study showed that how motivated and organized students are when they study greatly improves the accuracy of predictions.
3. Altabrawee *et al.* (2019) [3] compared Naïve Bayes, Decision Trees, and Artificial Neural Networks (ANN). They found that ANN was the most accurate (77.04%), but it was also the hardest to understand. They said that even simpler models, such as Decision Trees, can still help you figure out which traits are the most important.
4. Akter *et al.* (2025) [1] used Explainable AI (XAI) to use socio-economic and academic statistics to predict how well college students would do. The results showed that Random Forest was the most accurate (98.68%) at figuring out that attendance and economic background were important factors.

All of these studies show that Random Forest and other ensemble methods often work better than easier algorithms. However, the situation, the amount of data, and how easy it is to understand must be taken into account when choosing a model. But a lot of the studies that have been done so far use pretty big datasets. This leaves open the question of how these models work with smaller, institution-specific datasets, which is what this study is all about.

Statement of Problem

Educational institutions frequently encounter difficulties in identifying students who are at risk of underperformance until late in the academic cycle, which restricts the effectiveness of remedial interventions. Exam scores and teacher observations are the primary focus of current evaluation methods, which neglect critical engagement and behavioural factors. Subsequently, numerous students who are at risk are overlooked until their academic performance

has already declined considerably. This discrepancy emphasizes the necessity of data-driven methodologies that incorporate a variety of student information dimensions in order to generate precise and timely predictions.

Objectives of the Study

1. To look at student information like grades, attendance, and behavior to find trends in how they are doing.
2. To guess how well students will do by using Decision Tree, Random Forest, SVM, and KNN.
3. To test and compare how well these programs can predict the future.
4. To give teachers information they can use to help students who are at risk.
5. To encourage making decisions in schooling based on data.

Research Methodology

Data Collection

A set of data about 100 students was put together, showing a smaller sample in Table 1. The data included attendance rates, task submission rates, prior exam scores, and levels of participation. Students were marked as "Pass" or "Fail" by the goal variable.

Table 1: Sample Extract of Student Dataset

Student ID	Attendance (%)	Assignments	Previous Score	Participation	Performance (Target)
1	90	8	85	High	Pass
2	60	5	55	Low	Fail
3	80	7	78	Medium	Pass
4	50	4	45	Low	Fail

Data Processing

- Median (numerical) and mode (categorical) were used to fill in missing numbers.
- Levels of participation were recorded with numbers, (High = 2, Medium = 1, Low = 0).
- Min-Max scaling was used to make the features equal.
- The data was split into two groups: training (80%) and testing (20%).

Feature Selection

Through the use of correlation analysis and Recursive Feature Elimination (RFE), it was discovered that the most significant predictors were attended classes and scores on previous examinations. Through their subject knowledge, educators were able to corroborate this.

Model Development

Machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbours were implemented. In order to achieve the best possible performance, the default hyper parameters were optimized using grid search.

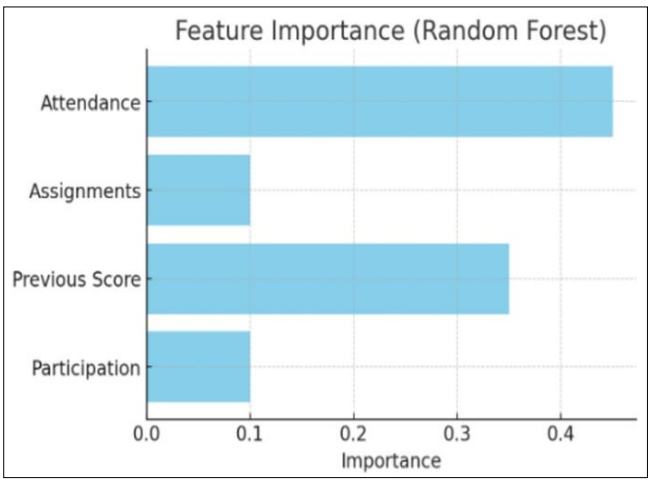
Evaluation Metrics

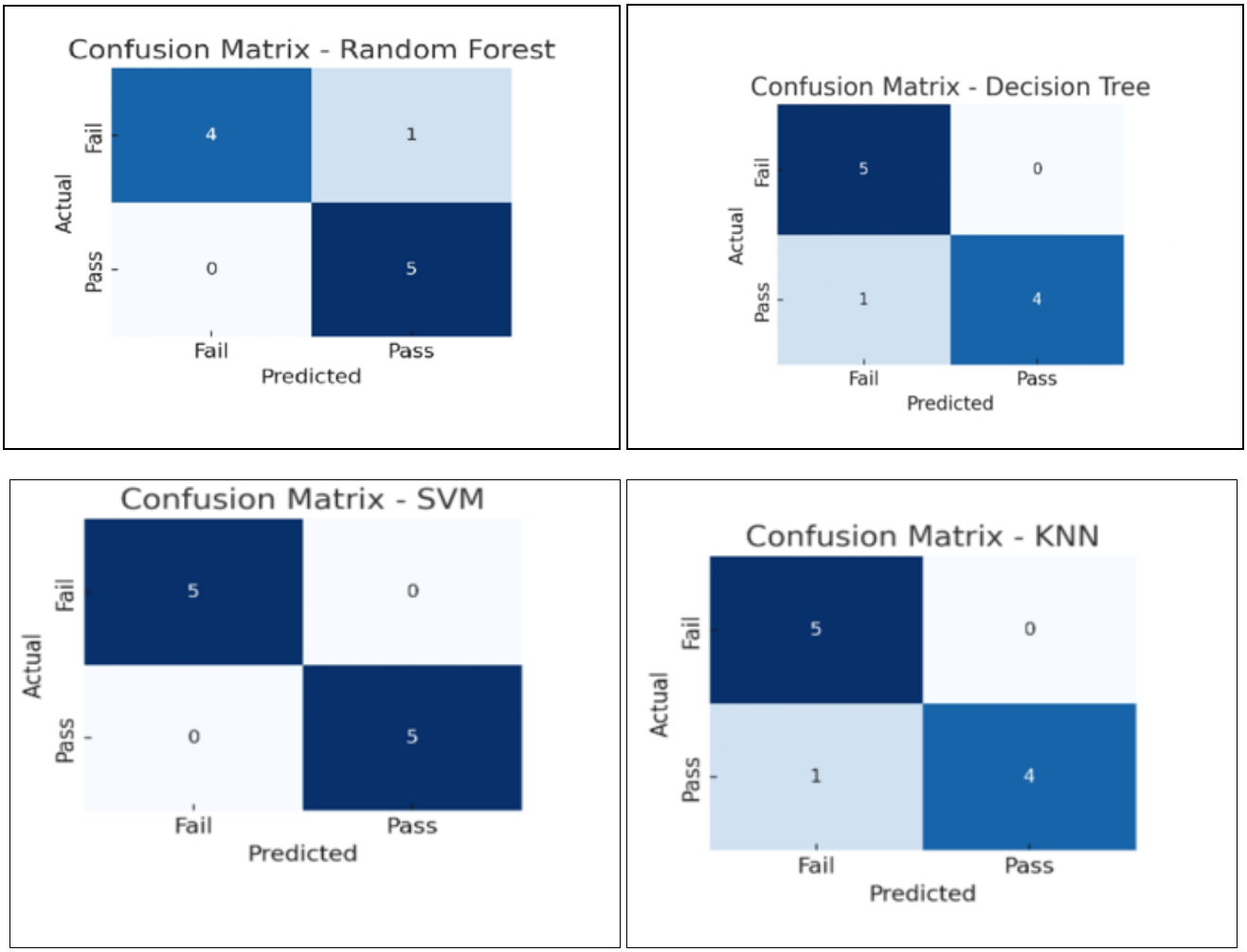
The Accuracy, Precision, Recall, F1-Score, and Confusion Matrices were utilized in order to gauge the performance of the models. In the case of tree-based models, the importance of features was investigated.

Results

Table 2: Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	85	0.86	0.83	0.84
Random Forest	90	0.91	0.89	0.90
SVM	88	0.87	0.86	0.86
KNN	82	0.83	0.80	0.81





Graphs: Feature Importance (Decision Tree / Random Forest) - Confusion Matrix for each model

Observations

- Random Forest regularly outperformed other models, reaching the best accuracy.
- Attendance and past scores were prominent characteristics, as illustrated in feature importance plots.
- Random Forest achieved the highest accuracy.
- The confusion matrices showed that all of the models predicted "Pass" cases more accurately than "Fail" cases, which indicates that there is a slight distinction between the classes.
- The performance of KNN was not particularly impressive, which suggests that it is not ideal for datasets that are both tiny and imbalanced.

Discussion

The findings demonstrate that ensemble approaches, such as Random Forest, are reliable and efficient when it comes to predicting academic success. Although Decision Trees are less accurate than other decision-making tools, they are nonetheless useful for educators who are trying to comprehend particular choice rules because of their interpretability.

Due to their reasonable performance, support vector machines (SVM) have the potential to be competitive; nevertheless, they may require larger datasets or parameters that are more finely calibrated. Given that KNN performed poorly in this investigation, it is clear that its performance is highly dependent on the size and distribution of the dataset.

The fact that this study only used a very limited dataset is one of its limitations, which may limit the extent to which the findings may be generalized. Furthermore, socio-economic and motivational variables were not included, which should have been incorporated in order to further increase the accuracy of the model.

Conclusion

The findings of this study indicate that machine learning has the potential to play a significant part in forecasting the academic achievement of students. Random Forest surfaced as the most successful model, with Support Vector Machines (SVM) and Decision Trees following closely behind. The most significant predictors were attendance and previous scores, which highlights the significance of maintaining regular involvement and having past knowledge that is already present.

By utilizing such models, educational institutions are able to identify kids who are at risk at an earlier stage, which enables them to implement individualized treatments. Not only does this increase the outcomes for students, but it also raises the overall quality of the academic experience.

Future Scope

1. Adding socioeconomic, motivational, and digital learning measures to make predictions that are more accurate.
2. Using deep learning models like ANN and LSTM to find learning trends over time.

3. Making predictive tools that work in real time so that teachers can keep an eye on things all the time.
4. Doing studies across institutions to make sure that results are true in different settings.

References

1. Akter B, Hosen MB, Ahmed S, Anannya M, Hossain, MF. An explainable AI-based approach for predicting undergraduate students' academic performance. Journal of Educational Data Mining,2025:15(7):3550. https://doi.org/10.3390/jedm150703550
2. Alamri A, Alzahrani A, Alzahrani A. Predicting student academic performance using machine learning algorithms. International Journal of Advanced Computer Science and Applications,2020:11(5):202–208. https://doi.org/10.14569/IJACSA.2020.0110530
3. Altabrawee H, Ali OA, Ajmi SQ. Predicting students' performance using machine learning techniques. Journal of University of Babylon, Pure and Applied Sciences,2019:27(1):194–201. https://doi.org/10.29116/jubaps.2019.27.1.23
4. Orji FA, Vassileva J. Machine learning approach for predicting students' academic performance and study strategies based on their motivation. Education and Information Technologies,2022:27(4):4871–4892. https://doi.org/10.1007/s10639-021-10788-2
5. Sandeepa AGR, Mohottala S. Evaluation of machine learning models in student academic performance prediction, 2025. *arXiv*. https://doi.org/10.48550/arXiv.2506.08047