International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 10 | Oct 2026 www.irjet.net p-ISSN: 2395-0072

Student Performance Prediction in Exams Using Machine Learning: A Survey-Based Study at Colleges of Navi Mumbai.

Dr. Siji Nambiar[1], Bindhya Karkera[2], Mr. Bignesh Rawal[3], Debarghya Majumder[4]

[1]Assistant Professor, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur 400614, Maharashtra, India.

[234]Student, Department of MCA, D. Y. Patil Deemed to be University, School of Humanities & Sciences, Belapur 400614, Maharashtra, India.

Abstract - *Student performance is affected by many academic and personal factors, such as study hours, attendance, assignments, previous performance, failures, class participation, sleep, study environment, and family support. This research focuses on using machine learning to classify student performance into three categories: Low, Average, and High. For this study, data was collected from 73 students using a Google Form survey. The collected data was analysed using four machine learning algorithms: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). The models were used to predict the performance category of students based on the collected factors. The main purpose of this research is to demonstrate how machine learning can be used to analyse student-related data and classify their academic performance. The study can help in understanding the factors associated with student performance and may support teachers and educational institutions in identifying students who require additional academic guidance and support. The results show that machine learning can be used to classify student performance and identify patterns in student data. This study shows that factors related to study habits, attendance, academic performance, and personal conditions can be useful for understanding student performance. The research can help teachers and educational institutions identify students who may need additional academic support.*

Keywords - Machine Learning, Student Performance, Classification, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM).

INTRODUCTION

Online education and online examinations have become common ways of teaching and evaluating students. During online learning, information such as attendance, study hours, assignment completion, previous marks, participation, and examination performance can provide useful information about a student's learning.

Collecting this information helps in understanding student performance and learning behaviour.
Student performance is affected by many different factors. These factors can be academic, behavioural, and personal. For example, study hours, attendance, assignments, previous performance, failures, participation, sleep, study environment, and family support may affect student performance. Finding and understanding all these factors manually can be difficult and time-consuming.

This research focuses on developing a machine learning-based approach for classifying student performance into Low, Average, and High categories using academic, behavioural, and learning-related factors.

The main motivation for this research is to use machine learning to better understand student performance. Early identification of students who may need academic support can help teachers provide suitable guidance. This study also aims to understand the factors related to student performance and apply machine learning techniques to educational data.

Objectives of the study

The main objectives of this research are:

- To collect student academic and behavioural data through a structured survey.
- To clean, preprocess, and encode the collected data for machine learning.
- To classify students into Low, Average, and High-performance categories.
- To implement Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) algorithms.
- To evaluate and compare the model's using accuracy, precision, recall, and F1-score.
- To demonstrate individual student performance prediction using machine learning.

LITERATURE REVIEW

Zhang et al. (2021) studied how machine learning can be used to predict student performance. They reviewed different methods used to collect student data, train machine learning models, and predict academic results. Their study showed that machine learning can help

teachers understand students' progress and identify students who need extra support.

Kotsiantis, Pierrakeas, and Pintelas (2004) studied student performance prediction in distance learning. They used student information and assignment marks to predict academic results. Their research showed that machine learning can help identify students who may face difficulties in their studies. It can also help teachers provide support at the right time.

Khan and Ghosh (2021) reviewed different studies related to student performance prediction in classrooms. They discussed the factors that affect students' academic results and how educational data mining can be used to understand student performance. Their study showed that student data can help researchers find patterns and better understand learning outcomes.

López-Zambrano, Lara Torralbo, and Romero (2021) reviewed previous research on predicting student performance using data mining techniques. They studied different methods and algorithms used to predict students' academic results. Their review discussed the use of methods such as Random Forest and Support Vector Machine (SVM). The study showed that student data can be useful for predicting learning outcomes and identifying students who may need academic support.

Based on these previous studies, the present research uses machine learning to classify students into Low, Average, and High-performance categories. Data was collected from 73 students through a Google Forms survey. The dataset includes study hours, attendance, assignment completion, previous performance, number of failures, participation, sleep, study environment, and family support.

Four machine learning algorithms Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) are used in this study. Python in Google Collab is used for data preprocessing, model training, prediction, and evaluation. The aim is to compare these algorithms and find which one performs best on the collected dataset.

METHODOLOGY

Research Workflow

The research follows a step-by-step process to classify student performance using machine learning. First, student data is collected through a Google Form. The collected data is then cleaned and converted into a suitable format for machine learning.

After that, the important features are selected and the dataset is divided into training and testing sets. Standardization is applied where required. Finally, different machine learning models are trained, evaluated, and used for student performance prediction.

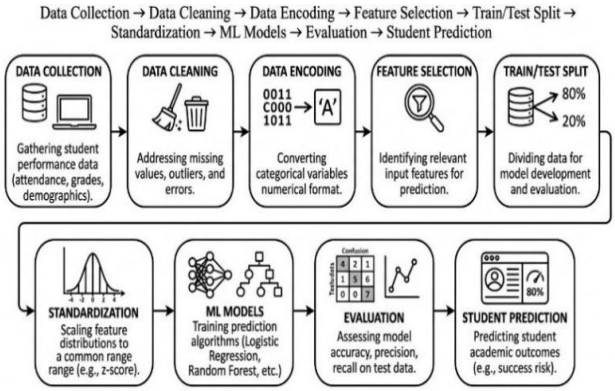


Figure of data workflow

Data Collection

The data for this research was collected through a Google Form survey. A total of 73 student responses were collected.

The survey was designed to collect information related to students' study habits, attendance, academic background, participation, and learning environment.

The following 14 features were collected and used as input variables for the machine learning models including study hours, coaching hours, attendance, assignment completion, previous academic performance, failed subjects, class participation, sleep duration, study environment, family support, social media usage, subject understanding, regular class attendance, and seeking help from teachers. Data was collected through a Google Forms survey of 73 students at D. Y. Patil College, Navi Mumbai, and analysed using multiple machine learning algorithms to predict academic performance.:

No.	Feature	Description
1	Study Hours	Number of hours spent studying
2	Study Hours at Outside Classes	Hours spent studying in outside classes or coaching
3	Attendance	Student attendance information
4	Assignment Completion	Level of assignment completion
5	Previous Performance	Student's previous

		academic performance
6	Failed Subjects	Number of subjects failed in the previous semester
7	Class Participation	Participation in class discussions and activities
8	Sleep Duration	Average number of hours of sleep per day
9	Study Environment	Student's rating of their study environment
10	Family Support	Level of family support for studies
11	Social media/Entertainment Time	Time spent on social media or entertainment
12	Subject Understanding	Student's level of understanding of subjects
13	Regular Class Attendance	Regularity of attending classes
14	Asking Teacher for Help	Frequency of asking teachers for help

Dataset Preparation

After collecting 73 student responses, the dataset was prepared for machine learning. The 14 input features were converted into numeric form so that they could be used by the machine learning algorithms.

The dataset was checked for missing and infinite values. The Previous Performance feature had 18 missing values. These missing values were replaced using the median value of the respective feature. Infinite values were also checked and handled during preprocessing.

After completing the preprocessing, the dataset contained 0 missing values and 0 infinite values. This confirmed that the dataset was ready for further analysis and model training.

Train/Test Split

The dataset was divided into training and testing data using the train_test_split method from Scikit-learn. 58 students (80%) training, while 15 students (20%) were use%) were used for d for testing.

A stratified split was applied to maintain the distribution of the three performance categories Low, Average, and High in both the training and testing datasets. A random state of 42 was used to make the data split reproducible.

Table of output representation

Dataset	Students	Features
Total dataset	73	14
Training set	58	14
Testing set	15	15

Standardization

Standardization is used to bring different numerical features to a similar scale. In this research, Standards scaler was used for Logistic Regression and SVM. This helps these models process the features properly when their values have different ranges.

Decision Tree and Random Forest were used without standardization because these models do not need feature scaling.

Machine Learning Models

Four machine learning models were used to classify students into three performance categories: Low, Average, and High. The models used in this research are Logistic Regression, Decision Tree, Random Forest, and SVM.

Logistic Regression

Logistic Regression is used for classification. In this research, it was used to predict whether a student's performance is Low, Average, or High.

Decision Tree

Decision Tree uses simple rules to make a prediction. It looks at the student's information and follows different decisions to classify the student's performance.

Random Forest

Random Forest uses many Decision Trees together to make a prediction. The results from the trees are combined to give the final performance category.

SVM

SVM is a classification method that separates different groups of data. In this research, it was used to separate students into Low, Average, and High-performance categories.

All four models were trained using the training data and tested using the testing data. Their results were then evaluated using accuracy, precision, recall, and F1-score.

Experimental Setup

The experiment was conducted using Python in Google Collab. The dataset was collected through a Google Form survey and initially stored in an Excel file. A total of 73 student responses were used in this study.

The dataset contained 14 input features related to students' study habits, attendance, academic performance, participation, sleep, study environment, family support, and other learning-related factors.

Before applying the machine learning models, the collected data was checked and prepared for analysis. Missing values were first checked and some values were handled manually in Excel.

The remaining missing values were handled during the Python preprocessing step by replacing them with the median value of the respective feature. After preprocessing, there were no missing or infinite values in the dataset.

Four machine learning algorithms were implemented: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). Standard scaler was applied to Logistic Regression and SVM, while Decision Tree and Random Forest were trained without feature scaling. The implementation used Python libraries including Pandas, NumPy, and Scikit-learn. Matplotlib & Confusion matrix was used for visualization where required.

The main experimental settings are summarized below:

Component	Details
Platform	Google Colab
Programming Language	Python
Dataset Source	Google Form
Dataset Format	Excel
Total Responses	73
Input Features	14
Target Classes	Low, Average, High
Training Data	58 students (80%)
Testing Data	15 students (20%)
Data Split	80:20
Missing Value Handling	Manual checking in Excel and median filling in Python
Algorithms	Logistic Regression, Decision Tree, Random Forest, SVM
Main Libraries	Pandas, NumPy, Scikit-learn
Evaluation Methods	Accuracy, Precision, Recall, F1-Score, Confusion Matrix
Visualization	Matplotlib

Data Analysis

Confusion Matrix Analysis

To further evaluate the performance of the machine learning models, a confusion matrix was used for each model. The confusion matrix compares the actual performance of the students with the predicted performance given by the model. In this study, the performance was divided into three classes: Low, Average, and High. The diagonal values in the confusion matrix represent the students who were correctly classified by the model.

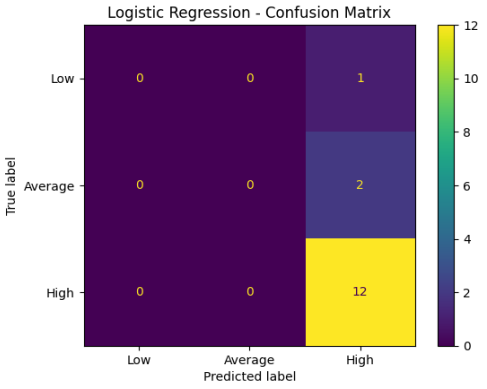


Figure 1: Confusion Matrix – Logistic Regression

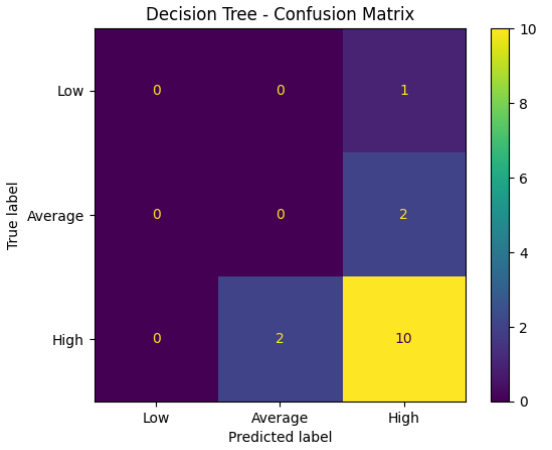


Figure 2: Confusion Matrix – Decision Tree

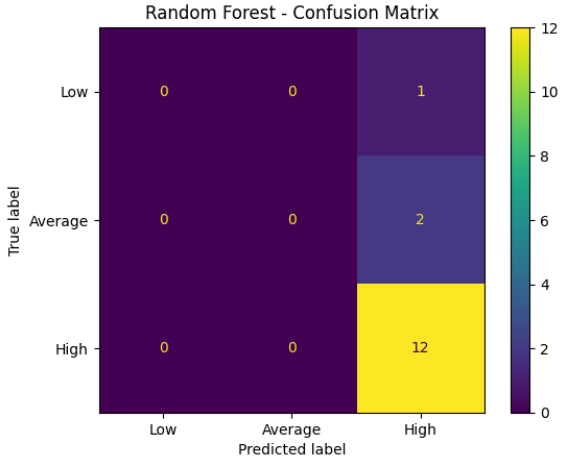


Figure 3: Confusion Matrix – Random Forest

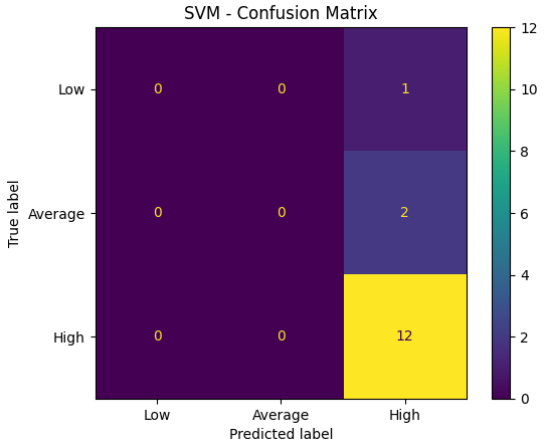


Figure 4: Confusion Matrix – SVM

The confusion matrices for Logistic Regression, Decision Tree, Random Forest, and SVM were generated using Scikit-learn and Matplotlib and are presented in the following figures.

Results and Evaluation

After completing the data preprocessing and training the machine learning models, the models were evaluated using the testing dataset. The testing dataset contained 15 student responses. Four evaluation measures were used: accuracy, precision, recall, and F1-score. These measures help to understand how well the model's classified students into the Low, Average, and High-performance categories.

The results obtained from the four models are presented in Table 1.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	80.00%	64.00%	80.00 %	71.11 %
Decision Tree	66.67%	61.54%	66.67 %	64.00 %
Random Forest	80.00%	64.00%	80.00 %	71.11 %
SVM	80.00%	64.00%	80.00 %	71.11 %

Logistic Regression, Random Forest, and SVM obtained an accuracy of 80% on the test dataset, while Decision Tree obtained 66.67%. The weighted F1-score for the first three models was 71.11%, compared with 64.00% for Decision Tree. Since the test set contains only 15 students and the dataset has an uneven distribution among performance categories, these results should be interpreted with caution.

Model Training

The models were trained using 58 student responses, which represented 80% of the complete dataset. During training, the models learned patterns between the 14 input features and the target performance categories. After training, the models were tested using the remaining 15 student responses.

A stratified 80:20 train/test split was used so that the performance categories were represented in both the training and testing datasets. The same testing data was used to evaluate all four models.

Model Evaluation

The trained models were evaluated using accuracy, precision, recall, and F1-score. These measures were calculated by comparing the predicted performance

categories with the actual categories in the testing dataset.

Accuracy shows the percentage of students that were classified correctly.

Precision shows how accurately the model predicted a particular performance category.

Recall shows how well the model identified students who actually belonged to a particular performance category.

F1-score combines precision and recall into one measure and provides a balanced view of classification performance.

The evaluation results are based on the 15 testing responses. Logistic Regression, Random Forest, and SVM produced the same values for the reported evaluation measures, while the Decision Tree produced different values. The confusion matrices provide additional information about the correct and incorrect classifications made by each model.

Individual Student Prediction

After training and evaluating the four machine learning models, the models were applied to the complete student dataset to generate individual performance predictions. For each student, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) predicted one of three performance categories: Low, Average, or High.

The individual prediction results were recorded for all 73 students along with their actual performance categories. These results allow the predicted performance of each student to be compared with their actual performance. The prediction table also helps identify cases where different machine learning algorithms produced different predictions for the same student.

The individual prediction results demonstrate the practical application of the proposed system, where academic and behavioural factors can be used to classify the expected performance category of an individual student.

It should be noted that these individual predictions are presented to demonstrate the application of the trained models. Model performance and generalization were evaluated separately using the held-out test dataset.

Small sample table representation of individual performance prediction data as below given:

Student No	Actual Performance	Logistic Regression	Decision Tree	Random Forest	SVM
1	high	high	High	high	High
2	low	low	low	low	low
3	high	high	Average	high	high
4	high	high	High	high	high

The predicted results were compared with the students' actual performance to check how well the models performed.

Feature Importance

the Random Forest model was used to find which features were more important for predicting student performance. Figure no. 3 shows the feature importance obtained from the Random Forest model.

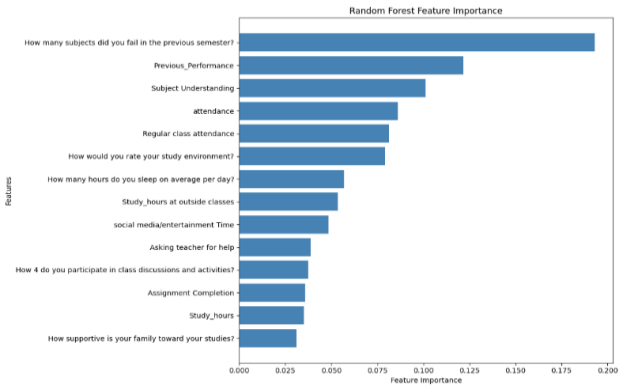
The results show that the number of subjects failed in the previous semester was the most important feature, with an importance value of 0.193225.

High importance

The next important features were Previous Performance (0.121552) and Subject Understanding (0.101250). Attendance (0.086146) and Regular Class Attendance (0.081454) also had relatively high importance.

Lower importance

Other features had lower importance values. Study Environment had an importance of 0.079165, followed by Sleep Hours (0.056969) and Study Hours at Outside Classes (0.053543). Social media/Entertainment Time (0.048562), Asking Teacher for Help (0.038722), and Class Participation (0.037343) had lower importance. Assignment Completion (0.035707), Study Hours (0.035108), and Family Support (0.031256) had the lowest importance values.



Graphical Representation

Interpretation and Discussion of Results

The study compared four machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), for predicting student performance as Low, Average, or High.

Comparison of Model Results

The results show that Logistic Regression, Random Forest, and SVM produced the same overall performance, with an accuracy of 80.00%, precision of 64.00%, recall of 80.00%, and F1-score of 71.11%.

The Decision Tree showed a lower performance, with an accuracy of 66.67%, precision of 61.54%, recall of 66.67%, and F1-score of 64.00%.

This shows that the three models, Logistic Regression, Random Forest, and SVM, gave similar overall results, while the Decision Tree produced different results.

Classification Performance

The models were used to classify students into three performance categories: Low, Average, and High. The overall accuracy of 80.00% for Logistic Regression, Random Forest, and SVM indicates that these models correctly classified a large proportion of the test students.

However, the precision value of 64.00% shows that some of the predicted categories did not match the actual categories. The recall value of 80.00% indicates that the models were able to correctly identify a relatively large proportion of the actual cases.

The Decision Tree achieved an accuracy of 66.67%, which was lower than the other three models. Its precision, recall, and F1-score were also lower. This suggests that the Decision Tree had more difficulty classifying the students correctly in this dataset.

Important Features

The Random Forest feature-importance results showed that some features contributed more to the model's predictions than others. The number of subjects failed in the previous semester had the highest importance value (0.193225). This was followed by Previous Performance (0.121552) and Subject Understanding (0.101250). Attendance (0.086146), Regular Class Attendance (0.081454), and Study Environment (0.079165) also received relatively higher importance values.

In comparison, Study Hours (0.035108) and Family Support (0.031256) had lower importance values. Therefore, in this Random Forest model, previous academic performance, subjects failed, subject understanding, and attendance-related features had higher relative importance than some of the other features.

Effect of the Small Dataset

The dataset contained 73 students, with 58 observations used for training and 15 observations used for testing. This is a relatively small dataset for machine learning. Because only 15 students were included in the test set, the performance measures can be sensitive to individual predictions.

A different train-test split could produce somewhat different results. Therefore, the findings should be interpreted carefully, and a larger dataset could be used in future research to obtain more reliable results.

Effect of Class Imbalance

The distribution of students across the Low, Average, and High categories can also affect model performance. If one category contains more students than another, the models have more examples from that category to learn from. This can make classification easier for the larger category and more difficult for categories with fewer observations.

For this reason, accuracy alone should not be used to evaluate the models. Precision, recall, and F1-score should also be considered when interpreting the results. The results of this study show that Logistic Regression, Random Forest, and SVM had the same overall evaluation scores, while the Decision Tree had lower scores on the given test data.

Key Findings

The main findings of this study are:

1. Logistic Regression, Random Forest, and SVM achieved the same accuracy of 80.00% on the test dataset.
2. Decision Tree achieved an accuracy of 66.67%, which was lower than the other three models.
3. Logistic Regression, Random Forest, and SVM also had the same precision of 64.00%, recall of 80.00%, and F1-score of 71.11%.
4. The Random Forest results showed that the number of subjects failed in the previous semester had the highest feature importance (0.193225).

5. Previous Performance (0.121552) and Subject Understanding (0.101250) were also among the features with higher importance.
6. Attendance (0.086146) and Regular Class Attendance (0.081454) also showed relatively higher importance.
7. The study shows that machine learning can be used to classify student performance into Low, Average, and High categories using academic and behavioural factors.
8. The results should be interpreted carefully because the dataset contained only 73 students, with 15 students in the test set, and the classes were imbalanced.

LIMITATIONS AND FUTURE RESEARCH

This study has some limitations. First, the dataset was small and included only 73 student responses. The data was collected through a survey, so the results depend on the information provided by the students.

The performance categories, Low, Average, and High, were based on the final-result information collected in the dataset. The number of students in each category was also not equal, which means the data was imbalanced. This may have affected the performance of the machine learning models.

Another limitation is that only 15 students were included in the test set. Because the test set was small, a few incorrect predictions could have a noticeable effect on the model results.

The results may also not apply to students from other colleges, schools, institutions, or different groups of students because the data was collected from a specific group.

In future research, a larger dataset can be used to get more reliable results. More academic and personal factors can also be included, such as exam marks, academic workload, learning resources, and other factors related to student performance.

CONCLUSION

This study implemented four machine-learning classification algorithms, namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), to classify student performance into Low, Average, and High categories.

The experimental results showed that Logistic Regression, Random Forest, and SVM achieved an accuracy of 80.00% on the test dataset, while the Decision Tree achieved 66.67% accuracy. The three models with 80.00% accuracy also had the same precision (64.00%), recall (80.00%), and F1-score (71.11%) in the evaluation.

The Random Forest feature-importance results showed that the number of subjects failed in the previous semester, Previous Performance, Subject Understanding, and attendance had relatively higher importance in the model.

Overall, the study shows that machine learning can be used to classify student performance using academic and behavioural factors. However, the results should be considered carefully because the study used a small dataset of 73 students, had an imbalanced class distribution, and used only 15 students for testing. Future research can use a larger and more diverse dataset and include additional academic and contextual factors to improve the analysis.

FUTURE SCOPE

There are several ways this study can be improved in the future. First, a larger dataset can be collected from different colleges to get more reliable results. More students from the Low, Average, and High-performance categories can also be included so that the classes are better balanced.

Future studies can include more academic factors, such as exam marks, academic workload, and other factors related to student performance. Different machine learning algorithms can also be tested and compared.

Cross-validation and hyperparameter tuning can be used to improve the model results. In the future, the model could also be developed into a web-based system where students or teachers can enter student information and get a performance prediction.

The system could also provide personalized academic suggestions based on the predicted performance. Finally, the model can be regularly tested with new student data to check whether its performance remains reliable over time.

REFERENCES

1. 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, Article 698490. https://doi.org/10.3389/fpsyg.2021.698490

2. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' performance in

distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411–426. https://doi.org/10.1080/08839510490442058

3. Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and Information Technologies, 26(1), 205–240. https://doi.org/10.1007/s10639-020-10230-3

4. López-Zambrano, J., Lara Torralbo, J. A., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. Psicothema, 33(3), 456–465. https://doi.org/10.7334/psicothema2021.62

5. R. Deshmukh, A. Kulkarni, A. Kumthekar, P. Kottur, and S. Raut, "Machine Learning Techniques for Predicting Student Performance," Mathematical Statistician and Engineering Applications, vol. 71, no. 4, pp. 168–189, 2022.
6. Student Performance Prediction Using Machine Learning is a 2021 conference paper by Priya S, Ankit T, and Divyansh D on applying educational data mining (EDM