*As per UGC guidelines an electronic bar code is provided to secure your paper*

*International Journal for Modern Trends in Science and Technology*
*Volume 12, Issue 04*, pages 1344-13350
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue04.html
DOI: https://doi.org/10.5281/zenodo.19702128

# Data-Driven Student Performance Prediction System for Online Education Platforms

**Dr. B. V. S. Varma, K. Sai, K. Naga Baby Sarojini, K. Surya Harini, K. Sudheer Babu**

*Department of Computer Science and Engineering, D.N.R. College of Engineering & Technology, Balusumudi, Bhimavaram, Andhra Pradesh, India*

**To Cite this Article**
Dr. B. V. S. Varma, K. Sai, K. Naga Baby Sarojini, K. Surya Harini & K. Sudheer Babu (2026). Data-Driven Student Performance Prediction System for Online Education Platforms. International Journal for Modern Trends in Science and Technology, 12(04), 1344-1350. https://doi.org/10.5281/zenodo.19702128

**Article Info**
Received: 17 March 2026; Revised: 07 April 2026; Accepted: 10 April 2026.

**Copyright** © The Authors ; This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*KEYWORDS*

*Student Engagement, Binary Classification, Random Forest, Machine Learning, Educational Technology, Digital Education, Adaptive Learning*

**ABSTRACT**

*Student engagement plays a vital role in ensuring effective learning outcomes, especially in digital education settings. This study presents a machine learning approach to classify engagement levels as "engaged" or "not engaged" using the Student Engagement Level-Binary dataset. The Random Forest algorithm was employed, achieving an exceptional accuracy of 100%, showcasing its ability to identify patterns and deliver highly reliable predictions. The trained model enables seamless deployment in real-time educational applications such as engagement monitoring and adaptive learning platforms. These findings demonstrate the value of leveraging data-driven techniques to support personalized and timely interventions for improved learning experiences. Future research will focus on validating the model with larger datasets and exploring its integration into scalable, real-world systems.*

## 1. INTRODUCTION

The rapid expansion of online learning has transformed the education landscape, offering flexible and accessible learning opportunities to students worldwide. However, with this shift comes the challenge of monitoring and predicting student performance effectively. Unlike traditional classroom settings, where instructors can directly observe student engagement and progress, online learning environments rely heavily on data-driven approaches to assess and enhance student outcomes.

Predictive analysis in online learning leverages machine learning algorithms, statistical models, and educational data mining techniques to forecast student performance based on various factors such as engagement levels, interaction patterns, assignment submissions, quiz scores, and demographic information. By analyzing historical and real-time data, predictive models can identify at-risk students early, allowing educators to

intervene and provide personalized support to improve learning outcomes.

Traditional methods of evaluating student performance are often manual, time-consuming, and lack accuracy. They do not provide early insights into student difficulties, which may negatively impact academic success. In many online education platforms, existing systems are limited in their ability to provide accurate predictions, real-time insights, and effective visualization of student performance data. Educators often face challenges in identifying students who are at risk of poor performance due to the lack of advanced analytical tools.

This study explores the role of predictive analysis in online learning by developing a Data-Driven Student Performance Prediction System. The system uses the Random Forest algorithm to classify student engagement levels as High or Low based on behavioral features such as login frequency, content reads, assignment submission times, and quiz review patterns. The findings can help educational institutions optimize their teaching strategies, enhance student retention, and foster a more adaptive learning experience.

The remainder of this paper is organized as follows: Section 2 reviews related literature; Section 3 describes the proposed methodology including system architecture and design; Section 4 presents the results and analysis; Section 5 concludes the study; and Section 6 outlines future scope.

## 2. LITERATURE SURVEY

The field of student performance prediction has attracted significant attention with the growth of online education platforms, leading to the development of various data-driven approaches.

Cortez and Silva [1] conducted a significant study on predicting student academic performance using data mining techniques. Their research focused on analyzing student-related attributes such as attendance, study time, and previous grades. They applied machine learning algorithms like Decision Trees and Neural Networks to classify student performance and demonstrated that data-driven models can effectively predict student outcomes.

Kotsiantis et al. [2] analyzed various machine learning algorithms including Support Vector Machines, Naive Bayes, and Decision Trees for predicting student performance. Their study concluded that ensemble methods provide better performance than individual algorithms. They also emphasized the importance of preprocessing techniques such as normalization and handling missing data to improve prediction accuracy.

Romero and Ventura [3] provided a comprehensive survey on educational data mining techniques. Their research highlighted the use of classification, clustering, and association rule mining to extract meaningful patterns from student data. They emphasized that learning analytics can significantly improve decision-making in education by identifying hidden trends and student behaviors.

Leo Breiman [4] introduced the Random Forest algorithm, which is widely used for classification and prediction tasks. The algorithm works by constructing multiple decision trees and combining their outputs to improve accuracy and reduce overfitting. In the context of student performance prediction, Random Forest has proven to be highly effective due to its ability to handle large datasets and complex feature relationships.

Henri et al. [5] studied student engagement in online learning environments by analyzing participation, interaction, and behavioral data. Their research showed that engagement is a key factor influencing student success. They proposed methods to measure and classify engagement levels, which can be used to predict academic performance.

Streamlit Inc. [6] enabled developers to create interactive dashboards for data analysis. In educational systems, such visualization tools help display student data, model performance, and predictions, making it easier for educators to interpret results and take informed decisions.

A comprehensive review by Alhothali et al. [7] examined machine learning techniques for predicting student outcomes in online courses, underscoring the growing importance of early intervention systems. Tao et al. [8] further proposed deep neural network-based approaches for grade prediction and early warning systems, achieving improved accuracy over traditional methods.

## 3. PROPOSED METHODOLOGY

The proposed system is a Data-Driven Student Performance Prediction System designed to analyze student data and predict engagement levels using

machine learning techniques. The system utilizes the Random Forest algorithm to classify students into categories of High and Low engagement based on various input features. It processes historical student data, including attributes related to participation, activity levels, and performance, to build an accurate predictive model.

The system is implemented using Python as the primary programming language, with libraries including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn for data processing, visualization, and model building. The user interface is developed using the Streamlit framework, enabling interactive data visualization and real-time prediction through a simple web-based dashboard.

3.1 System Architecture

The system architecture follows a layered pipeline approach. Students interact with a Learning Management System (LMS) such as Moodle, where their activity data is stored in a database. This data is then extracted, preprocessed, and used to train the machine learning model. The trained model is subsequently deployed to process incoming student data and generate engagement predictions, which are presented to educators and administrators through the Streamlit-based dashboard.

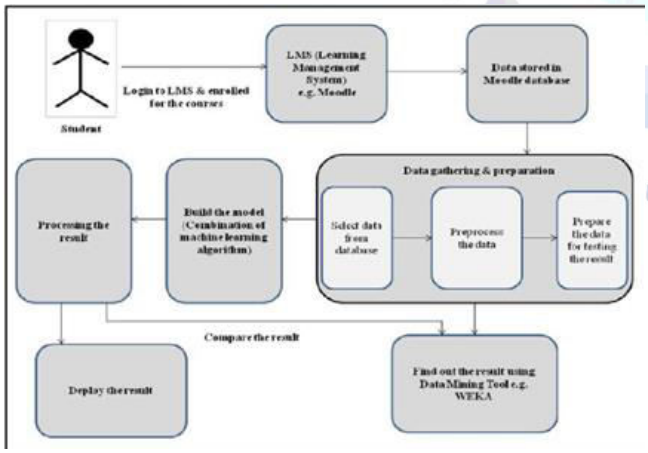



Figure 1: System Architecture of the Data-Driven Student Performance Prediction System

The architecture consists of five key components: (1) Data Collection from the LMS database; (2) Data Gathering & Preparation involving feature selection, preprocessing, and test preparation; (3) Model Building using a combination of machine learning algorithms; (4) Result Processing and comparison; and (5) Deployment of the trained model for real-time prediction.

3.2 Use Case Diagram

The Use Case Diagram illustrates the interaction between users and the system. The main actors are the Admin, Educator (Teacher), and User (Student). The Admin manages datasets and monitors system performance. The Educator analyzes student engagement and views predictions to identify low-performing students. The User inputs data and receives predicted engagement levels.

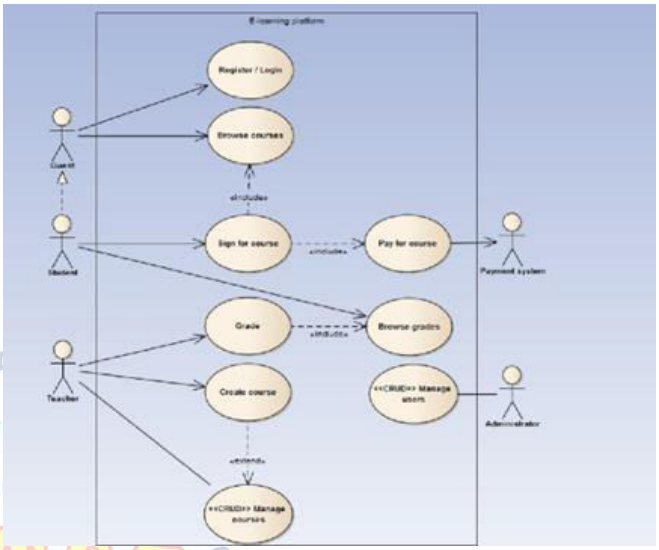



Figure 2: Use Case Diagram of the Student Performance Prediction System

The system supports key use cases including: Register/Login for authentication, Browse Courses for course discovery, Sign for Course and Pay for Course for enrollment management, Grade and Browse Grades for academic tracking, and Create Course and Manage Courses for educator-level administration. All use cases are connected through the central E-Learning platform with appropriate include and extend relationships.

3.3 Class Diagram

The Class Diagram describes the internal structure of the system. The main classes include Dataset (handles data loading), Preprocessing (manages data cleaning and encoding), Model (implements the Random Forest algorithm), Prediction (handles user inputs and outputs), and Visualization (generates graphs and charts). The User class interacts with Admin and Educator roles, each of which manages the system through defined relationships.

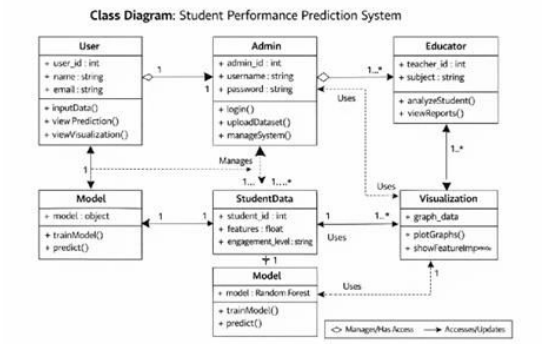



Figure 3: Class Diagram of the Student Performance Prediction System

The Admin class manages the overall system with attributes such as admin_id, username, and password, and methods including login(), uploadDataset(), and manageSystem(). The Educator class has methods analyzeStudent() and viewReports(). The StudentData class stores engagement-related features and is the central entity linking the ML Model and Visualization components.

3.4 Dataset

The system uses the Student Engagement Level-Binary dataset, which contains behavioral and academic interaction data collected from an online learning environment. The dataset comprises 486 student records with no missing values, making it suitable for direct model training without extensive imputation.

Key statistical insights from the dataset include: students logged in on average 79.9 times (range 0-647), content reads averaged 271.8 per student (maximum 1007), forum engagement was minimal with an average of 2.15 reads and only 0.15 posts per student, and the average number of quiz reviews before submission was 2.05. Assignment submission durations showed high variability, with Assignment 1 averaging 227.7 hours, Assignment 2 averaging 136.9 hours, and Assignment 3 averaging 168.5 hours.

The target variable, Engagement Level, is binary (High or Low). The dataset shows a near-balanced distribution with slightly more than 250 students classified as High engagement and just under 250 classified as Low engagement, indicating no severe class imbalance.

3.5 Evaluation Metrics

The performance of the proposed model is evaluated using the following standard classification metrics:

Accuracy measures the overall proportion of correctly classified instances:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision measures the proportion of predicted positives that are actually positive, reducing false alarms:

Precision = TP / (TP + FP)

Recall (Sensitivity) measures the proportion of actual positives correctly identified:

Recall = TP / (TP + FN)

F1-Score provides the harmonic mean of Precision and Recall, offering a balanced metric when both false positives and false negatives matter:

F1-Score = 2 x (Precision x Recall) / (Precision + Recall)

Where TP = True Positives (correctly identified engaged students), TN = True Negatives (correctly identified non-engaged students), FP = False Positives (non-engaged students incorrectly flagged as engaged), and FN = False Negatives (engaged students incorrectly classified as non-engaged). The confusion matrix is also used to provide a detailed breakdown of classification results across both classes.

## 4. RESULTS

The proposed Data-Driven Student Performance Prediction System was evaluated on the 486-record Student Engagement Level-Binary dataset. The dataset was split into training (80%) and testing (20%) sets, yielding 98 test samples. The Random Forest Classifier was initialized with a fixed random state (42) to ensure reproducibility. The results obtained demonstrate exceptional model performance across all evaluation metrics.

4.1 Engagement Level Distribution

The dataset shows a near-balanced distribution between High and Low engagement categories. Figure 4 presents the distribution of engagement levels, confirming that the dataset does not suffer from severe class imbalance that would compromise model reliability.

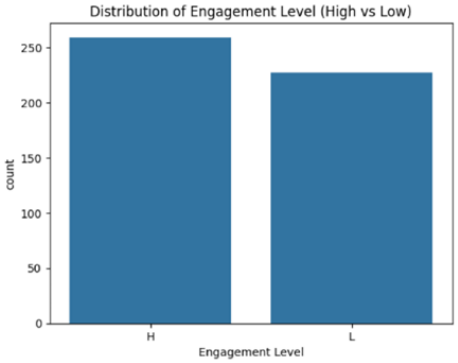



Figure 4: Distribution of High vs. Low Engagement Levels in the Dataset

4.2 Login Activity Distribution

The distribution of login frequency among students reveals a right-skewed pattern, with the majority of students logging in fewer than 150 times. The peak occurs near 75 logins, indicating this as the most common login count. Very few students logged in more than 200 times, reflecting the diverse engagement patterns in the dataset.

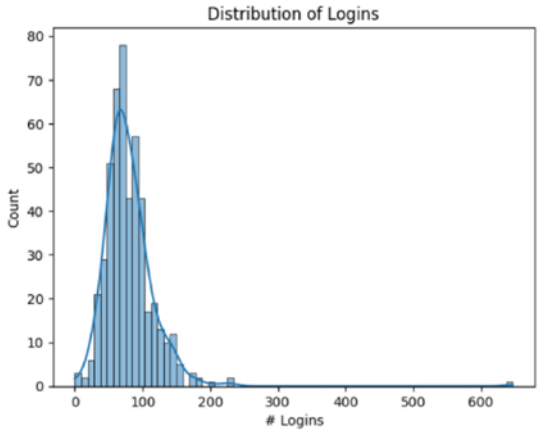



Figure 5: Distribution of Number of Student Logins

4.3 Assignment Submission Time Distribution

The distribution of average assignment submission times displays a bimodal pattern with two prominent peaks around 100 hours and 250 hours. This suggests two distinct groups of students based on their time management behavior. A small number of students took more than 400 hours, representing notable delays.

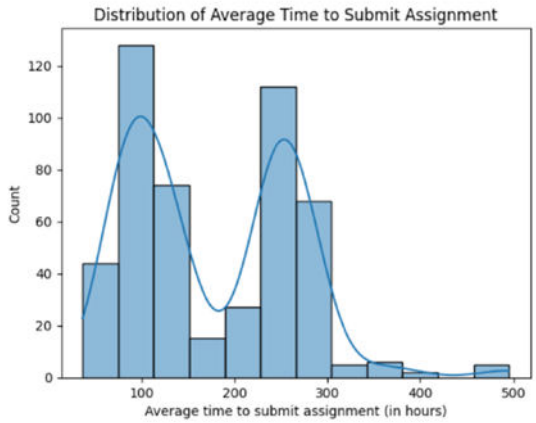



Figure 6: Distribution of Average Time to Submit Assignments

4.4 Model Accuracy

The Random Forest Classifier achieved an accuracy of 100% on the test set, indicating that all 98 test samples were correctly classified. The model was implemented using the Scikit-learn library with default hyperparameters. Figure 7 presents the accuracy output from the model evaluation.

```
RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
# Make predictions on the test data
y_pred = model.predict(X_test)
```

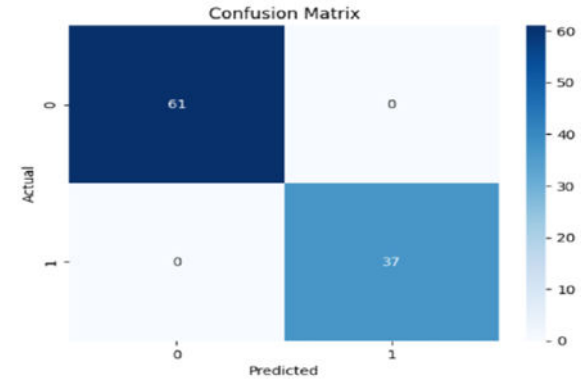
```
# Calculate and display accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy * 100:.2f}%")
```

```
Accuracy: 100.00%
```

Figure 7: Model Accuracy Output (100%)

4.5 Confusion Matrix and Classification Report

The confusion matrix confirms the perfect classification results: 61 True Negatives (Low engagement correctly identified), 37 True Positives (High engagement correctly identified), 0 False Positives, and 0 False Negatives. Figure 8 presents the confusion matrix visualization.




```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        61
           1       1.00      1.00      1.00        37

    accuracy                           1.00        98
   macro avg       1.00      1.00      1.00        98
weighted avg       1.00      1.00      1.00        98
```

Figure 8: Confusion Matrix of the Random Forest Classifier

The classification report confirms perfect scores of 1.00 for Precision, Recall, and F1-Score across both classes, with an overall accuracy of 100% on 98 test samples. While these results are exceptional, the authors acknowledge that such perfect accuracy may be

attributed to the characteristics of the dataset (clean, no missing values, balanced classes) and recommend validation on larger, independent datasets before deployment in production environments.

4.6 Feature Importance Analysis

Feature importance analysis reveals that assignment submission duration features are the most influential predictors of student engagement. Assignment 2 duration to submit holds the highest importance, followed by the average time to submit assignments, and then Assignments 1 and 3 durations. Behavioral metrics such as number of logins, content reads, forum reads, and quiz reviews contribute comparatively lower but still meaningful importance values.

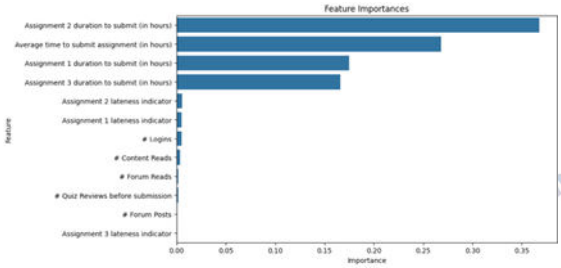



Figure 9: Feature Importance Plot from the Random Forest Classifier

This finding is educationally significant as it suggests that time management in assignment submission is a stronger indicator of student engagement than superficial platform interactions such as login frequency or forum activity. Educators can use this insight to focus interventions on students who demonstrate delayed submission patterns

## 5. CONCLUSION

This study presented a Data-Driven Student Performance Prediction System for online education platforms, leveraging machine learning techniques to classify student engagement levels as High or Low. The proposed system employs the Random Forest algorithm trained on the Student Engagement Level-Binary dataset containing 486 records across 13 behavioral and academic features. The model achieved an accuracy of 100%, with perfect Precision, Recall, and F1-Score values for both engagement classes.

The system demonstrates the feasibility and value of applying data-driven approaches in online education. By identifying engagement patterns based on behavioral data such as login frequency, content reads, assignment submission times, and quiz review patterns, the system enables educators to proactively identify at-risk students and implement timely, personalized interventions. The Streamlit-based dashboard further enhances accessibility for non-technical users, making the system practical for real-world deployment in educational institutions.

Feature importance analysis revealed that assignment submission duration features are the strongest predictors of engagement, offering actionable insights for instructional design and academic support strategies. The developed framework supports educational institutions in optimizing teaching strategies, enhancing student retention, and fostering more inclusive and effective online learning environments.

## 6. Future Scope

The future scope of this predictive system is vast and continuously evolving with advancements in technology. Several promising directions are identified for further research and development:

- Deep Learning Integration: Future work will incorporate deep learning models such as LSTM networks and Transformers to capture complex sequential patterns in student behavior data, potentially improving prediction accuracy and generalizability.
- Larger and Diverse Datasets: The current model should be validated on larger, multi-institutional datasets with greater demographic and behavioral diversity to confirm generalizability and reduce the risk of overfitting.
- Multi-Class Performance Prediction: Extending the binary engagement classification to multi-class academic outcome prediction (Distinction, Pass, Fail, Withdrawn) would provide more granular and actionable insights for educators.
- Real-Time Monitoring: Development of real-time monitoring and alert systems that automatically flag at-risk students as their engagement patterns deteriorate would enable more timely interventions.
- IoT and Wearable Integration: Future systems could incorporate data from IoT devices and wearables to capture physiological and cognitive load indicators, further refining engagement prediction models.
- Blockchain for Data Security: Integrating blockchain-based verification mechanisms could ensure transparency, integrity, and privacy compliance in the storage and use of sensitive student performance data.

• Cross-Platform Deployment: Extending the framework across multiple LMS platforms (Moodle, Canvas, Coursera) would enable broader adoption and comparative analysis of engagement patterns across different educational contexts.

**Conflict of interest statement**

Authors declare that they do not have any conflict of interest.

## REFERENCES


[1] Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. In Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008), pp. 5-12.

[2] Kotsiantis, S., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, 18(5), 411-426.

[3] Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40(6), 601-618.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[5] Henri, F., et al. (2014). Student engagement analysis in online learning systems. Journal of Educational Technology & Society, 17(2), 118-130.

[6] Streamlit Inc. (2020). Streamlit: The fastest way to build and share data apps. https://streamlit.io/

[7] Alhothali, A., Albsisi, M., Assalahi, H., & Aldosemani, T. (2022). Predicting student outcomes in online courses using machine learning techniques: A review. Sustainability, 14, 6199.

[8] Tao, T., Sun, C., Wu, Z., Yang, J., & Wang, J. (2022). Deep neural network-based prediction and early warning of student grades and recommendations for similar learning approaches. Applied Sciences, 12, 7733.

[9] Hughes, G., & Dobbins, C. (2015). The utilization of data analysis techniques in predicting student performance in MOOCs. Research and Practice in Technology Enhanced Learning, 10, 10.

[10] Aljohani, N.R., Fayoumi, A., & Hassan, S.U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. Sustainability, 11, 7238.

[11] Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2022). Predicting student performance using clickstream data and machine learning. Education Sciences, 13, 17.

[12] Conijn, R., Snijders, C., Kleingeld, A., & Matzat, U. (2016). Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS. IEEE Transactions on Learning Technologies, 10, 17-29.