# Mining Student Interaction Data to Predict Academic Performance and Dropout Risks

**[1]Dr.Bokare Madhav M.,[2]Mr.Suryawanshi Amol V**

[1]Associate Professor,[1]Assistant Professor
[1]Department of Computer Science,
[1]SSBES' Institute of Technology and Management, Nanded, India

***Abstract*** **-** Higher education institutions now have serious concerns about student retention and general academic performance. By analysinganalyzing activity logs from Learning Management Systems (LMS) and records from institutional databases, the current study investigates how educational data mining might be used to assess student behavioural trends. The investigation identifies significant engagement patterns that support academic advancement and early indicators of dropout risk using machine learning methods, including Random Forests, Support Vector Machines (SVMs), and Gradient Boosting. The findings show a substantial correlation between early-semester indicators and final academic outcomes, with a predicted accuracy of more than 85%. These indicators include the number of logins, involvement in discussion forums, timely assignment submission, and quiz performance. This research highlights the importance of predictive analytics as a mechanism for enabling proactive academic support, improving student success, and enhancing institutional retention strategies.

***Index Terms*** **– SVM, Random Forest, Predication, LMS, EDM.**

## I. INTRODUCTION

The way educational institutions monitor and improve student learning has changed in recent years due to the growing use of Learning Management Systems (LMS). These digital platforms continuously record large amounts of data, including attendance records, discussion participation, assessment results, and general contact logs. Such data can be a valuable resource for understanding student learning behaviour, academic progress, and emerging risk indicators when analysed systematically. By identifying students who are performing poorly or exhibiting early indicators of disengagement, teachers can take appropriate action to improve learning outcomes and student retention.

Despite significant advancements in educational data mining, many current research methods concentrate on a small number of variables, usually attendance or grades, which limits the accuracy and dependability of predictive models. To better understand student learning patterns, there is an increasing need for more inclusive frameworks that incorporate academic, behavioural, and interaction-based characteristics. To better identify students who are at risk, the current study presents an improved machine-learning-driven prediction system that examines a variety of LMS indicators. In addition to improving predictive performance, the model is intended to assist institutions in implementing data-guided interventions to raise success rates and lower dropout rates.

## II. PROBLEM STATEMENT

Identifying students who are likely to do poorly or leave before the situation gets urgent is a common challenge for universities and other higher education institutions. Ongoing behavioural and engagement patterns are not captured by traditional evaluation approaches, which mostly rely on manual observation and recurring examinations. Thus, to effectively predict academic risk and analyze student interaction data in real time, a dynamic, automated analytical framework is needed. In addition to strengthening institutional efforts to increase retention and academic achievement, such a system would enable prompt intervention strategies.

## III. OBJECTIVES

- To analyze student interaction data from LMS and institutional databases.
- To identify behavioural factors influencing academic performance and dropout.
- To develop predictive machine learning models for risk detection.
- To evaluate the performance of different algorithms, including Random Forest, SVM, and Gradient Boosting.
- To design a risk-alert framework for academic intervention and retention improvement.

## IV. Research Design or Flow of System

**Load the dataset**

- Reads student_interaction_dataset.csv.
- Separates features (student activity) and label (pass/fail).

**Preprocessing**

- Scales all feature values using StandardScaler to improve model performance.

**Train-Test Split**

- Splits data into:
  - 80% for training
  - 20% for testing

**Model Training**

- Trains a Random Forest Classifier to predict if a student will pass or fail.

**Predictions**

- Uses a trained model to predict pass/fail on test data.

**Evaluation**

- Calculates:
  - Accuracy
  - Classification report (precision, recall, F1-score)
  - Confusion matrix

**Visualisations**

Creates 6 graphs:

1. Histograms → Distribution of time spent & login count
2. Scatter plot → Time spent vs quiz score
3. Bar chart → Login comparison: pass vs fail
4. Correlation heatmap → Relationship between features
5. Confusion matrix heatmap
6. Feature importance → Which activity affects the prediction the most

**Final Prediction Column**

Adds a new column predicted_dropout with model results (0 = fail risk, 1 = pass).

## V. Literature Review

The use of learning analytics and educational data mining (EDM) to predict student dropout risk and academic success is growing. Many studies focus on using machine learning models to analyse student records, behavioural logs, and demographic data. Yağcı [1] achieved prediction accuracy of 70%–75% by using Random Forests, SVMs, Logistic Regression, and related algorithms to predict undergraduate test scores based on academic and demographic characteristics. Similarly, Roslan and Chen [2] conducted a comprehensive review of 58 EDM research papers (2015–2021). They discovered that, while categorisation models such as Decision Trees and Random Forests are often used, academic performance and demographic characteristics are the most popular predictors.

Researchers in [3], who used data from 291 university students to identify academic achievement early, made a substantial contribution. Their study demonstrated that early indicators can accurately predict results by proposing a segmentation paradigm based on early-semester activity trends and academic conduct. Numerous EDM publications lack datasets on behaviour or interactions, according to thorough evaluations. Only a small percentage of the assessments examined LMS logs, forum involvement, or clickstream interaction data as predicted criteria, according to a comprehensive evaluation [4]. This restriction makes it more difficult to spot early warning signs in kids' learning styles. Additionally, comparison studies reveal conflicting findings about the advantages of machine learning models over conventional statistical techniques. In many instances, generalised linear regression performed better than sophisticated models such as Random Forests and Decision Trees, according to a recent study based on data from [5], especially when predictions were generated using smaller academic datasets.

In conclusion, the research highlights the following gaps: (1) a strong emphasis on academic and demographic variables rather than comprehensive behavioural interaction data; (2) a narrow focus on early-stage dropout prediction; and (3) an inadequate assessment of model efficacy across various algorithm groups. To enable early and precise prediction of academic risks and dropout patterns, these findings underscore the need for research that combines data from multiple sources (LMS logs, submissions, time-spent indicators) with machine learning models.

Examines the evolution of prediction accuracy in a university context, demonstrating that dropout risk differs among various student groups and that the impact of variables like GPA increases over time [6]. focuses on clarity and useful, actionable predictions while proposing a dependable machine learning model (Random Forest plus Decision Trees) to predict academic success [7].

## VI. SAMPLING PLAN

**Step-by-Step Research Process**

1. **Data Collection**

Login activity of students, scores on quizzes, attendance records, participation in forums, and submissions for assignments.

2. **Data Preprocessing**
   - Cleaning, standardization, and managing absent data
   - Extraction of features (metrics of engagement)
3. **Model Development**

   **Apply ML algorithms:**

   Random Forest

   SVM

   Gradient Boosting

   Split the dataset into training and testing sets (e.g., 70:30 ratio)
4. **Model Evaluation**
5. **Result Analysis & Risk Alert System**

**Tools Used:** Python, SQL Database

## VII. SYSTEM ALGORITHM AND WORKFLOW

**Algorithm: Student Performance & Dropout Prediction**

Import dataset D from the CSV file
Obtain feature matrix X and label vector y
Utilise StandardScaler:
X_scaled ← scale(X)
Divide dataset:
    (X_train, X_test, y_train, y_test) ← split_train_test (X_scaled, y)
Set up Random Forest Classifier:
    model ← RandomForestClassifier()
Train model:
    model.fit(X_train, y_train)
Make a Prediction on test data:
    y_pred ← model.predict(X_test)
Assess model:
    Calculate accuracy and classification summary
    Calculate the confusion matrix
Assess dropout risk for the complete dataset:
    dropout_labels ← model.predict(X_scaled)
Create visual representations:
    a. Histogram of essential characteristics
    b. Scatter graph of duration spent vs exam score
    c. Bar graph of login counts for dropouts compared to non-dropouts.
    d. Matrix of correlations
    e. Confusion matrix visualization
    f. importance plot of Features
Provide return accuracy, y_pred, dropout_labels, and visualizations.

**Input**: Student interaction dataset D
**Output**: Predicted performance labels and dropout risk

## VIII. RESULTS AND DISCUSSION

The following figure shows the system's output.

```
C:\Users\desir\AppData\Local\Programs\Python\Python311\python.exe C:\Users\desir\PycharmProjects\MiningProject\main.py

Model Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        77
           1       1.00      1.00      1.00        23

    accuracy                           1.00       100
   macro avg       1.00      1.00      1.00       100
weighted avg       1.00      1.00      1.00       100
```

Figure 1 Output

**Classification Report**

Table 1 Classification Report

| Metric | Description |
|---|---|
| **Precision** | Percentage of correct positive predictions out of all positive predictions. Here, 1.00 means all predictions for both pass (1) and fail (0) are correct. |
| **Recall** | Percentage of actual positives correctly identified. 1.00 means the model caught all actual pass/fail students correctly. |
| **F1-score** | Harmonic mean of precision and recall. 1.00 means a perfect balance between precision and recall. |
| **Support** | Number of actual instances in each class (77 fail, 23 pass). |
| **Accuracy** | Overall correctness of the model predictions (100%). |
| **Macro average** | Average of metrics across both classes without weighting. |
| **Weighted average** | Average weighted by the number of instances per class. |

**Interpretation:**

- The model achieved **perfect accuracy (1.0)** on the test set.
- It correctly classified all students into **pass (1)** or **fail (0)** categories.

No misclassifications occurred in this sample.

**Sample Dropout Predictions:**

Table 2 Sample Dropout Predictions

| Column | Meaning |
|---|---|
| time_spent | Hours spent on the platform. |
| login_count | Number of times the student logged in. |
| predicted_dropout | Model prediction: 0 = Fail risk, 1 = Pass |

**Interpretation of Sample:**

- Example 0 → Student spent 112 hrs, logged in 44 times → Predicted 0 (Fail risk)
- Example 2 → Student spent 280 hrs, logged in 27 times → Predicted 1 (Pass)

Shows that the model considers multiple features (time_spent, login_count, assignments, quiz_score, etc.) to predict dropout risk, not just a single metric.

**Graphical Representations:**

The dataset was visualised using various graphs to understand overall student behaviour and identify patterns influencing dropout rates. The correlation heatmap shows a strong relationship between factors such as time spent and login frequency, helping identify the main predictors. The distribution graphs for each attribute highlight differences in student engagement across the dataset, emphasizing the contrast between high- and low-performing students. The bar chart showing the number of dropouts versus non-dropouts illustrates the class distribution in the data. The feature-importance graph from the Random Forest model clearly indicates the primary factors that significantly affect dropout prediction. Together, these visualisations provide a quick, straightforward understanding of the dataset and support the model's decision-making process.

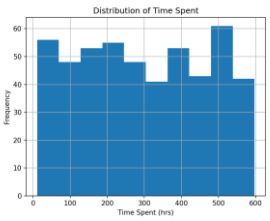


Figure 2: Distribution of Time Spent

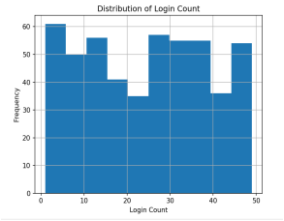


Figure 3: Distribution of Login Count

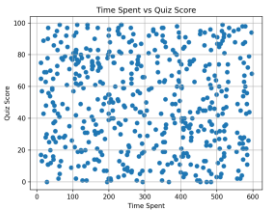


Figure 4: Time Spent vs Quiz Score

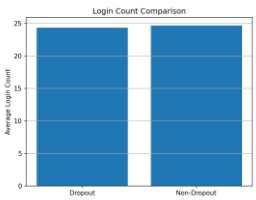


Figure 5: Login Count Comparison

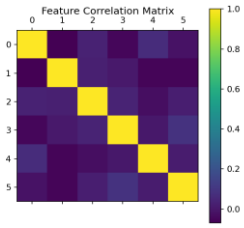


Figure 6 Feature Correlation Matrix

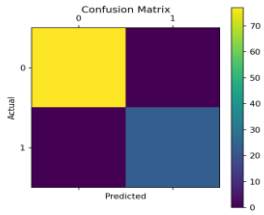


Figure 7: Confusion Matrix

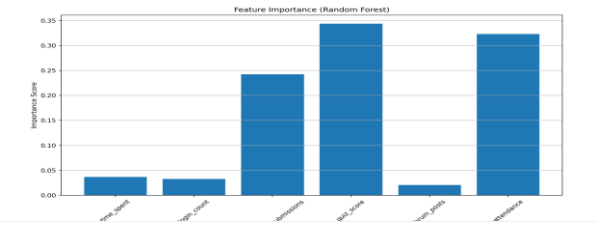


Figure 8 Feature Importance (RF)

Table 3: Meaning of all graphs

| Fig | Visualisation | Meaning |
|---|---|---|
| **1** | Histograms | Spread and allocation of student engagement |
| **2** | Scatter Plot | Correlation between time invested and quiz results |
| **3** | Bar Chart | Comparison of logins between passing and failing students |
| **4** | Correlation Matrix | Interconnection among all attributes |
| **5** | Confusion Matrix | Effectiveness of model predictions |

## IX. CONCLUSION

According to recent research, analysing LMS participation significantly enhances the ability to identify academic threats. However, very few studies integrate behavioural, performance, and social interaction data to create a comprehensive predictive model. To sum up, this approach lays the groundwork for a more comprehensive, proactive approach to academic hazards in educational establishments.

## REFERENCES

[1] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 11, 2022.
[2] M. H. Roslan and C. J. Chen, "Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 5, 2022.
[3] "Early detection of student degree-level academic performance using educational data mining," *PLOS ONE*, 2023.
[4] "Educational data mining, student academic performance prediction: overview of review studies 2013-2021," *Journal of e-Learning and Knowledge Society*, 2022.
[5] "Performance prediction using educational data mining techniques," *Springer*, 2025.
[6] Dominik Glandorf, Hye Rin Lee, et al., "Temporal and Between-Group Variability in College Dropout Prediction" (2024)
[7] Gul, M. N., Abbasi, W., & Wani, M. Y., "Revolutionizing educational decision making: a robust machine learning mechanism for predicting student performance" (2025).