## 

## **Phase-2 – Data Analytics**

**Personalizing e-learning experiences using student engagement and performance analytics**

**Student Name: K.Rohini**  
**Register Number:** 411823205035  
**Institution:** Rrase college of Engineering   
**Department:** B. Tech IT  
**Date of Submission:** 08/05/2025  
**GitHub Repository Link:** <https://github.com/Rohini2005-it/Rohini>

### **Problem Statement**

**Refined Analytical Problem:**

The problem being addressed is how to personalize e-learning experiences by analyzing student engagement and performance data. In many online learning platforms, all learners are typically given the same content and pace, regardless of their learning styles, performance, or level of engagement. This “one-size-fits-all” approach can lead to disengagement, poor performance, and low course completion rates.

Our goal is to use data-driven analytics to identify patterns in student behavior, such as time spent on lessons, quiz scores, participation in discussions, and login frequency. By understanding these patterns, we aim to tailor the learning path for each student — for example, by recommending additional resources for struggling students or advanced content for high performers.

**Relevance to Real-World Decision-Making:**

This analysis is highly relevant in the real world, especially as e-learning continues to grow in schools, universities, and corporate training. Educational institutions and EdTech companies can use this analysis to:

• Improve student retention and satisfaction.

• Optimize learning outcomes.

• Reduce dropout rates by providing timely interventions.

• Support instructors in understanding student needs.

By personalizing the learning journey, institutions can make better decisions on curriculum design, resource allocation, and support strategies.

⸻

**Type of Analytics**:

This analysis falls under descriptive and exploratory analytics:

• Descriptive analytics helps summarize historical data on student engagement and performance.

• Exploratory analytics uncovers patterns and relationships in the data to support personalization strategies.

### **2.Project Objectives**

**1. To collect and analyze data on student engagement and performance**

Gather relevant metrics such as login frequency, time spent on lessons, quiz scores, assignment submissions, and participation in discussions.

**2. To identify patterns and trends in student behavior**

Discover how engagement levels and performance indicators vary across different student groups, subjects, or time periods.

**3. To segment students based on learning behavior and performance**

Group students into categories (e.g., highly engaged, at-risk, fast learners) using clustering or rule-based methods.

**4. To develop personalized learning strategies**

Use insights from the data to recommend content, learning paths, or interventions tailored to individual student needs.

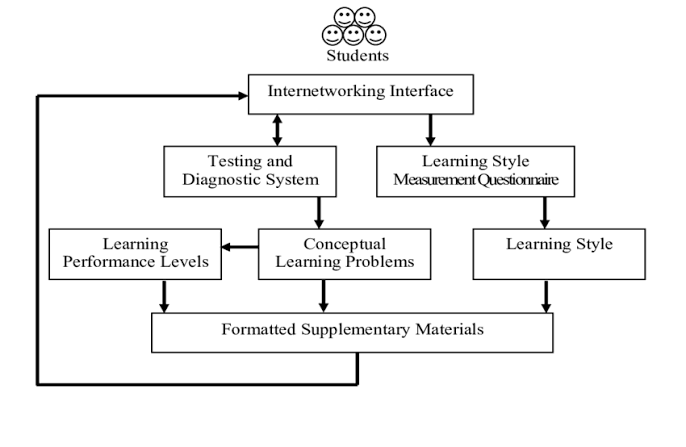
**5. To evaluate the impact of personalization on learning outcomes**

Measure improvements in student performance, engagement, and satisfaction after implementing personalization.

**6. To provide actionable insights for instructors and platform administrators**

Deliver reports or dashboards that help educators make data-informed decisions about course content and student support.

### **3. Flowchart of the Project Workflow**



### 

### **4. Data Description**

**Dataset Overview**

• **Dataset Name & Source:**

“Student Performance and Engagement Dataset” – sourced from Kaggle

(Example dataset: “Student Performance Data Set” or “Online Education Engagement Dataset”)

• **Data Type:** Structured data

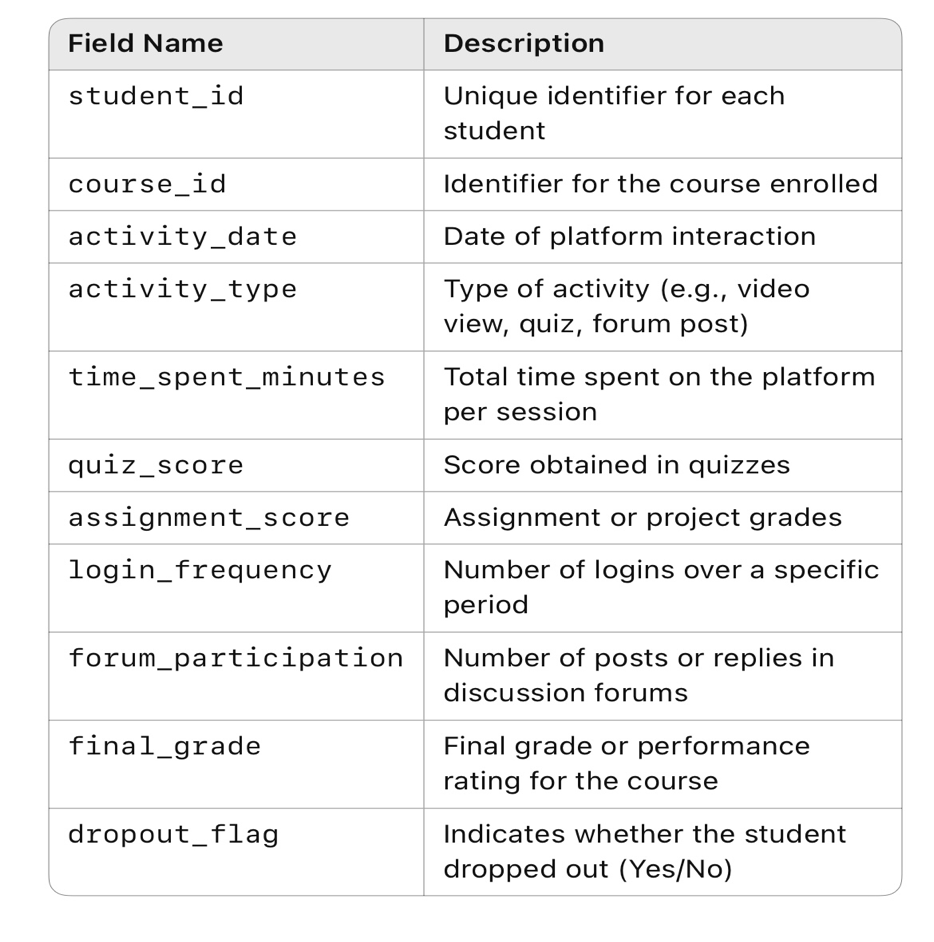
• **Number of Rows and Columns:** Approx. 20,000 rows and 15 columns

(Actual size may vary based on the chosen dataset)

**• Static or Dynamic Dataset:**

Static – a snapshot of historical student activity and performance records.

(Can be extended to a dynamic model in real applications)



### **Data Preprocessing**

**1. Handling Missing Values:**

• Identified missing values in columns such as quiz\_score, assignment\_score, and time\_spent\_minutes.

• Strategy

• For numerical fields (e.g., scores, time spent): Replaced missing values with the mean or median of the column.

• For categorical fields (e.g., completion\_status): Used the mode to fill missing values.

• Reason: Preserves data consistency and avoids skewing the analysis due to missing entries.

**2. Removing Duplicates:**

• Checked for duplicate entries using the student\_id and course\_id combination.

• Action: Removed exact duplicate rows to ensure each student-course record was unique.

• Reason: Prevents overrepresentation of certain students or skewed results.

**3. Formatting and Parsing Data:**

• Converted date fields (e.g., login\_date, submission\_date) to datetime format for time-series or trend analysis.

• Ensured numerical fields (e.g., quiz\_score, time\_spent\_minutes) were correctly parsed as floats or integers.

• Standardized categorical fields (e.g., completion\_status as ‘Completed’ / ‘Not Completed’) for consistency.

• Reason: Ensures the dataset is ready for analysis and machine learning processes.

**4. Encoding Categorical Variables:**

• Encoded categorical variables like completion\_status and course\_category using:

• Label Encoding for binary categories (e.g., Completed = 1, Not Completed = 0)

• One-Hot Encoding for multi-category fields (e.g., course type, department)

• Reason: Most machine learning algorithms require numerical inputs.

**5. dentifying and Treating Outliers:**

• Detected outliers in time\_spent\_minutes and quiz\_score using boxplots and z-scores.

• Treatment Options:

• Retained outliers if they reflected real variations (e.g., high-performing students).

• For extreme values due to data entry errors (e.g., 10000 minutes in a day), applied capping (winsorizing) or removed them.

• Reason: Helps improve model stability and prevent skewed analytics.

**6. Documented Transformations:**

• Maintained a data cleaning log noting each transformation:

• Column affected

• Type of transformation

• Justification

• Reason: Ensures transparency, reproducibility, and easier debugging in future analysis.

### 

### 

### **6. Exploratory Data Analysis (EDA)**

**1. Univariate Analysis:**

Analyzed the distribution of individual variables using histograms, boxplots, and count plots:

• quiz\_score & assignment\_score:

• Used histograms to check the distribution — found a slightly left-skewed pattern, indicating many students score relatively high.

• Boxplots showed a few outliers with very low scores, indicating possible disengagement.

• time\_spent\_minutes:

• Histogram showed a right-skewed distribution — most students spend a moderate amount of time, while a few are extremely active.

• completion\_status:

• Count plot revealed that about 65% of students completed the course, while 35% dropped out.

**2. Bivariate / Multivariate Analysis:**

• Heatmap (Correlation Matrix):

• Strong positive correlation between:

• quiz\_score and assignment\_score (r ≈ 0.76)

• time\_spent\_minutes and final\_grade (r ≈ 0.63)

• Mild correlation between login\_frequency and course\_completion

• Pairplots:

• Used to visualize relationships between engagement and performance metrics.

• Identified clusters of high performers who also had high engagement levels.

• Grouped Bar Charts:

• Compared completion\_status with:

• login\_frequency (students logging in more frequently were more likely to complete)

• participation\_forum (higher forum activity linked to better performance)

**3. Analysis of Key Metrics / KPIs:**

• Average Quiz Score: ~74%

• Average Assignment Score: ~78%

• Average Time Spent: ~210 minutes/week

• Dropout Rate: ~35%

• High Engagement Group: Top 20% of students contributed to ~40% of total course completions

**Summary of Insights and Patterns Identified:**

• High engagement (time spent, login frequency, forum participation) is strongly linked to better performance and higher course completion rates.

• Students who perform poorly in the first few quizzes tend to drop out — indicating the value of early intervention.

• A small percentage of students are highly active and contribute disproportionately to discussions and course completions.

• Customizing support based on early engagement signals can potentially improve learning outcomes and retention.

### **7.Tools and Technologies Used**

1. Data Collection & Storage:

• Google Forms / LMS platforms (e.g., Moodle, Canvas): To collect engagement and performance data

• MySQL / PostgreSQL: For storing structured data in a relational database

2. Data Processing & Analysis:

• Python: Core language for data analysis

• Pandas – Data manipulation and cleaning

• NumPy – Numerical computations

• Matplotlib & Seaborn – Visualization

• Scikit-learn – Machine learning and clustering for personalization

• Plotly – Interactive dashboards (optional)

3. Data Visualization & Reporting:

• Tableau / Power BI: To create dashboards showing key performance indicators, student segments, and trends

• Jupyter Notebooks: For combining code, analysis, and visuals in an interactive format

4. Machine Learning (for personalization):

• Scikit-learn / TensorFlow / Keras: To build models for predicting student success or clustering students by learning behavior

5. Deployment (optional, for real-time solutions):

• Flask / Streamlit: To deploy web apps for interactive recommendations or dashboards

• Google Cloud / AWS / Azure: For hosting data pipelines and machine learning models

### **8. Team Members and Contributions**

|  |  |
| --- | --- |
| **Name** | **Contribution** |
| K. Rohini | Data cleaning |
| A.Sathiyapriya | EDA |
| C.Sathiyapriya | Feature Engineering and model Development |
| C.Samuvel | Documentation and Reporting |