## Project Summary

| | |
|---|---|
| Batch details | **PGP DSG JUL 25** |
| Team members | **Mr. Akash N S**<br>**Ms. Aparna Abhilash**<br>**Mr. Deepak Arjun K**<br>**Mr. Karuturi Harshith Mani Sriram**<br>**Mr. Shashwath P** |
| Domain of Project | **Ed-Tech Analytics** |
| Proposed project title | **Student Course Completion Prediction** |
| Group Number | **4** |
| Team Leader | **Mr. Deepak Arjun K** |
| Mentor Name | **Mr. Mohit Sahu** |

Date: 19/11/2025

Signature of the Mentor

Signature of the Team Leader

# Table of Contents


| Sl NO | Topic | Page No |
|---|---|---|
| 1 | Overview | 1 |
| 2 | Business problem goals | 1 |
| 3 | Topic survey in brief | 6 |
| 4 | Critical assessment of topic survey | 7 |
| 5 | Methodology to be followed | 7 |
| 6 | References | 10 |

# Project Details

## OVERVIEW:

The purpose of this project is to develop and design an effective predictive model for an EdTech organization that aims to forecast whether a student will complete a course based on their demographic attributes, engagement levels, learning behaviors, and interaction patterns on the platform.

With the rapid growth of online learning, understanding and predicting student performance has become crucial. This project leverages extensive historical student data to identify completion patterns and generate insights that help improve student retention, engagement, and overall learning outcomes.

## BUSINESS PROBLEM STATEMENT (GOALS):

### 1. Business Problem Understanding:

- Student retention is one of the most significant challenges in the online education industry. High dropout rates lead to reduced course completion percentages, revenue losses, and poor learner satisfaction.
- By accurately predicting whether a student is at risk of not completing a course, EdTech platforms can take timely intervention measures, provide personalized support, and improve the overall learning experience.
- Factors such as study hours, quiz performance, device access, learning style, mentorship interaction, and internet connectivity play a vital role in determining student success.
- Understanding these patterns and predicting completion likelihood enables proactive academic guidance, reduces dropout rates, and strengthens the platform's learning ecosystem.

### 2. Business Objective:

To build a robust machine learning model using the platform's student learning and engagement data to predict whether a student will complete a course.

- Identify students at high risk of non-completion

- Enable proactive academic and technical support
- Optimize mentorship and reminder strategies
- Improve course design based on engagement patterns
- Enhance overall student satisfaction and performance

This predictive capability empowers the EdTech platform to make data-driven decisions that directly contribute to increased course completion rates and improved learner success.

## 3. Approach:

**Understanding the Problem**

- Identify key factors that influence whether a student completes a course.
- Study engagement patterns such as study hours, quiz scores, attendance, and video-watching behavior.

**Data Exploration**

- Analyze distributions of major features to understand student behavior.
- Compare engagement differences between completed and non-completed students.
- Identify the strongest indicators of dropout.

**Data Preparation**

- Remove non-essential fields like Student_ID and Name.
- Encode categorical variables such as device type, learning style, and course domain.
- Standardize numerical values like study hours and quiz scores.

**Model Development**

- Compare performance to select the most reliable model.
- Train multiple classification models such as Logistic Regression, Random Forest, Decision Tree and boosting algorithms.

**Model Evaluation**

- Use accuracy, precision, recall, F1-score, and ROC-AUC to measure effectiveness.
- Identify which features contribute most to prediction.

**Insights & Intervention Strategy**

- Highlight behavioral patterns of at-risk students.
- Provide actionable insights for early intervention to improve student retention.

**Data Description:**

The data set includes information about:

- **Student Profile Information:** contains basic student information that aids in understanding learner demographics, including age, gender, education level, employment status, and city.
- **Course Enrolment & Performance:** Includes course details (Course ID, Name, Level, Duration) and performance metrics like assignments submitted/missed, quiz attempts, quiz scores, project grade, and completion status, allowing evaluation of academic progress.
- **Engagement & Activity Behaviour:** Tracks how actively students use the platform through login frequency, session duration, time spent, video completion rate, rewatch count, discussion participation, and peer interaction.
- **Technical & Usage Factors:** Captures device type and internet quality to analyse whether technical conditions affect learning behaviour or performance.
- **Financial & Enrolment Details:** Includes payment mode, fee paid, discount used, final payment amount, and enrolment date, supporting financial or trend-based analysis.
- **Support & Satisfaction Metrics:** contains support tickets raised and satisfaction rating, reflecting student experience and platform effectiveness.

**Data Size:**

- Number of Columns : 40
- Number of Rows : 1,00,000
- Total Number of Records : 1,00,000

**Data Dictionary:**

1. **Student_ID –** Unique identifier assigned to each student.

2. **Name –** Full name of the student.
3. **Gender –** Gender of the student (Male/Female/Other).
4. **Age –** Age of the student in years.
5. **Education_Level –** Highest education level completed by the student.
6. **Employment_Status –** Employment status of the student.
7. **City –** City where the student resides**.**
8. **Device_Type –** Type of device used by the student.
9. **Internet_Connection_Quality –** Quality of the student's internet connection.
10. **Course_ID –** Unique identifier of the course.
11. **Course_Name –** Name of the course enrolled.
12. **Category –** Category/subject area of the course.
13. **Course_Level –** Difficulty level of the course.
14. **Course_Duration_Days –** Duration of the course in days.
15. **Instructor_Rating –** Rating of the course instructor.
16. **Login_Frequency –** Number of login sessions by the student.
17. **Average_Session_Duration_Min** – Average duration (in minutes) that the student spends per session on the platform.
18. **Video_Completion_Rate** – Percentage of course videos completed by the student.
19. **Discussion_Participation** – Count or level of participation in course discussion forums.

20. **Time_Spent_Hours** – Total number of hours spent on the course/content.
21. **Days_Since_Last_Login** – Number of days since the student last logged into the platform.
22. **Notifications_Checked** – Number of notifications viewed or clicked by the student.
23. **Peer_Interaction_Score** – Score indicating interaction with peers (group work, chats, forums).
24. **Assignments_Submitted** – Number of assignments successfully submitted.
25. **Assignments_Missed** – Number of assignments missed by the student.
26. **Quiz_Attempts** – Number of quiz attempts made by the student.
27. **Quiz_Score_Avg** – Average quiz score obtained.
28. **Project_Grade** – Grade of the final project.
29. **Progress_Percentage** – Percentage of course completion.
30. **Rewatch_Count** – Number of times lessons/videos were replayed.
31. **Enrollment_Date** – Date when the student enrolled.
32. **Payment_Mode** – Mode of payment used by the student.
33. **Fee_Paid** – Total fee paid for the course.
34. **Discount_Used** – Amount of discount applied.
35. **Payment_Amount** – Final amount paid after discount.
36. **App_Usage_Percentage** – Percentage of time the student uses the mobile app vs. other devices.
37. **Reminder_Emails_Clicked** – Number of reminder emails clicked/opened.
38. **Support_Tickets_Raised** – Number of support tickets raised by the student.
39. **Satisfaction_Rating** – Student's satisfaction score for the course.

40. **Completed (Target)** – Indicates whether the student completed the course (Yes/No or 1/0).

### 4. Conclusion:

This project successfully develops a predictive model that identifies whether a student is likely to complete a course based on their engagement, demographics, learning behavior, and support interactions. By leveraging these insights, EdTech platforms can intervene early, offer personalized support, and enhance overall student retention and learning outcomes. The model provides a valuable foundation for data-driven decision-making, ultimately helping improve course effectiveness and learner success.

## TOPIC SURVEY IN BRIEF

### 1. Problem understanding:

Online learning platforms experience high dropout rates, which affect student performance and institutional success. Many students fail to complete courses due to low engagement, irregular participation, poor time management, and lack of personalized academic support. Understanding these causes is essential to improving completion rates.

### 2. Current solution to the problem:

Presently, institutions rely on simple tracking dashboards and automated reminders. While they provide basic progress updates, they do not accurately detect at-risk students early. These generic systems fail to adapt to individual learning patterns, making interventions less effective.

### 3. Proposed solution to the problem:

Using the dataset, a predictive model can be developed to identify students who are likely to drop out. Features such as login frequency, assignment submission rate, time spent on learning materials, and quiz scores help forecast completion likelihood. This allows educators to take proactive, personalized actions.

### 4. Reference to the problem:

The dataset is directly relevant because it includes key behavioral and performance indicators that influence course completion. By analyzing these patterns, institutions can deploy targeted interventions, reduce dropout rates, and enhance learning outcomes using data-driven strategies.

## CRITICAL ASSESSMENT OF TOPIC SURVEY

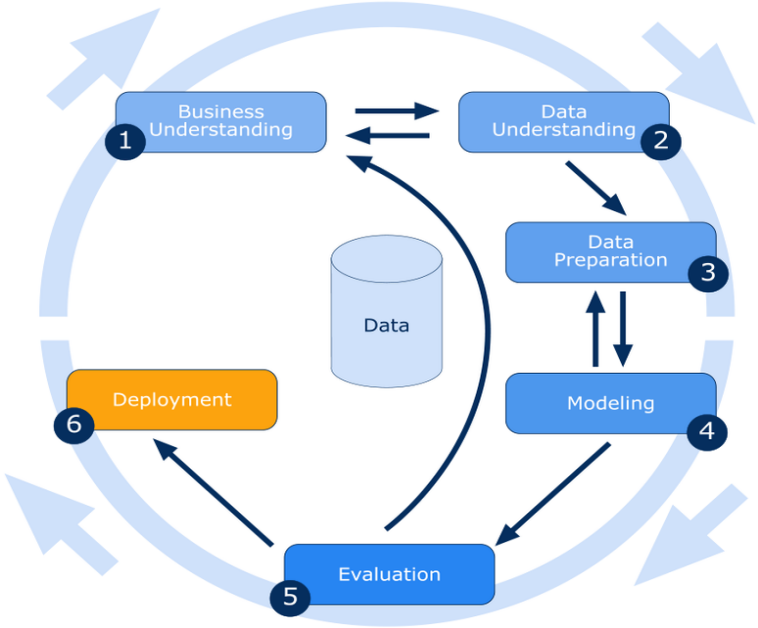
### 1. Key Areas and Gaps Identified:

The survey highlights significant gaps in current student monitoring systems, mainly the inability to predict dropouts early. While engagement and performance data are included, psychological and motivational factors are missing. Addressing these gaps can improve prediction accuracy.

### 2. Key Gaps the Project Aims to Solve:

The project aims to solve the major gap of early identification of at-risk students. By leveraging predictive analytics, institutions can implement timely interventions, offer personalized support, and enhance student retention in online courses.

## METHODOLOGY TO BE FOLLOWED



## 1. Business Understanding (Data Analysis, cleaning/ Preprocessing):

The pre-processing of the dataset before performing ML functions involves the following:

a. **Descriptive Analytics:**
Descriptive statistics summarize the core characteristics of the dataset, offering insights into its structure through measures like central tendency and variability. These statistics, combined with graphical analyses, lay the foundation for quantitative exploration, revealing the shape, spread, and overall distribution of the data.

b. **Inferential Analysis:**
Inferential methods validate insights derived from descriptive analysis, leveraging appropriate statistical tests to confirm the significance of observed patterns or relationships.

c. **Treating Outliers:**
Outliers in numerical columns are identified and analysed for their impact. Methods like the Interquartile Range (IQR) or other suitable techniques are applied to address them effectively.

d. **Treating Missing Values:**
Null or missing values in the dataset are handled using appropriate imputation strategies, ensuring data completeness and consistency.

e. **Encoding Categorical Variables:**
Machine learning models require numerical inputs. Categorical variables are encoded into numeric formats using suitable techniques, or dropped if deemed unnecessary for the modelling process, ensuring compatibility with mathematical computations.

f. **Dropping Unnecessary Columns:**
Columns that do not contribute meaningfully to model performance or hold little relevance are removed to optimize the dataset for analysis.

## 2. Data Understanding (Exploratory Data Analysis):

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Bar plot, Box plot, Scatter plot and many more using Univariate, Bivariate and Multivariate Analysis.

## 3. Data Preparation:

a. **Scaling:** It helps to normalize the data within a particular range and as well as in speeding up the calculations in an algorithm.
b. **Train and Test Split of the Data:** The data is split into train and test in required ratio.

## 4. Model Building :

We will try to fit/train and test with below ML models and compare the performances

1. Logistic Regression
2. Naive Bayes
3. KNN Classifier
4. Decision Tree
5. Random Forest
6. Bagging Classifier
7. Boosting Classifier

## 5. Model Evaluation:

Model Evaluation: Below metrics are used to evaluate the multi classification models performance.

1. Accuracy
2. Precision
3. Recall
4. F1-score

5. Confusion Matrix
6. RoC/AuC Score

## REFERENCES:


- https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success
- https://www.kaggle.com/datasets/nisargpatel344/student-course-completion-prediction-dataset
- https://www.sciencedirect.com/science/article/pii/S2211949323000170
- https://ieeexplore.ieee.org/document/11011419
- https://www.researchgate.net/publication/383069478_Predictive_Analysis_of_Online_Course_Completion_Key_Insights_and_Practical_Implications


## Notes For Project Team

*Sample Reference for Datasets (to be filled by team and mentor)*

| | |
|---|---|
| Original owner of data | UCI Machine Learning Repository |
| Data set information | 100,000 rows x 40 columns |
| Any past relevant articles using the dataset | NA |
| Reference | Yes |
| Link to web page | https://ieeexplore.ieee.org/document/11011419 |

**********