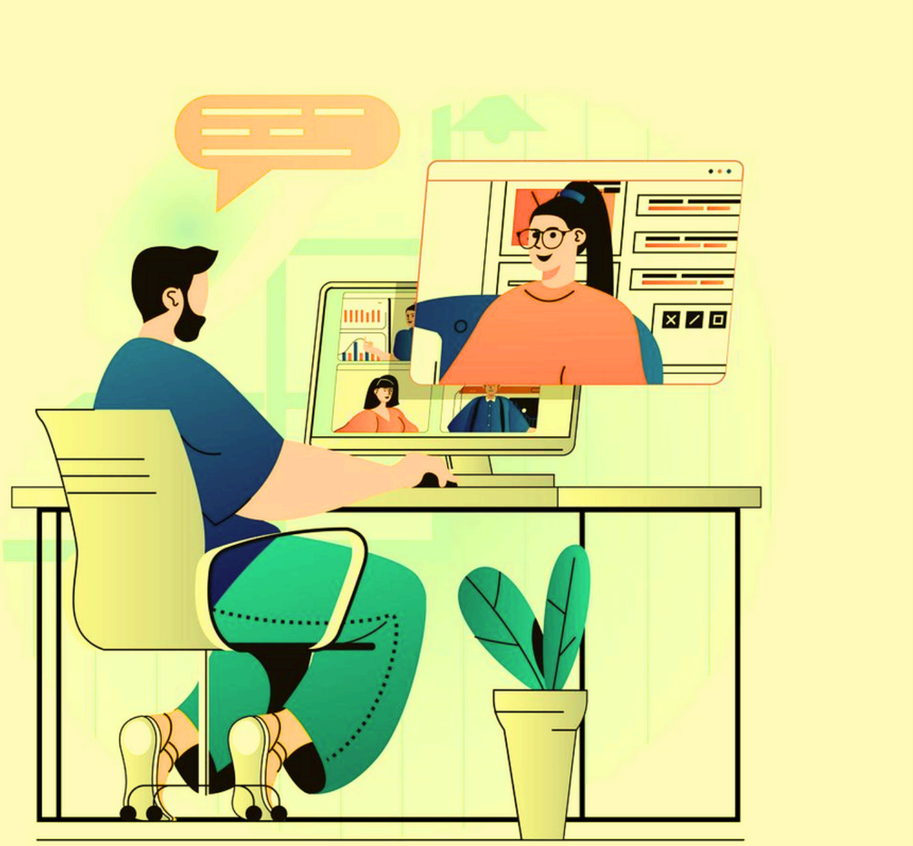
Online Course Engagement Prediction



Objective

The primary objective of this project is to identify key factors influencing course completion on an online learning platform by analyzing user engagement metrics, and to develop predictive models to forecast course completion.

Lifecycle of Analysis

01

Data Collection and Cleaning

02

Exploratory Data Analysis

03

Data Preparation

04

Modelling and Evaluation

About Dataset

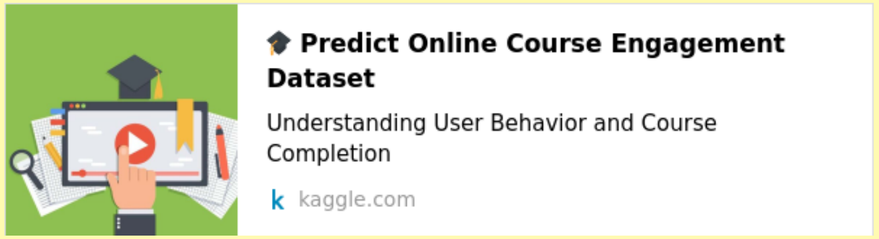
This dataset is designed to analyze factors influencing course completion on an online course platform. It contains a total of **9,000** instances, each providing insights into user engagement and course interaction.

The dataset includes **eight** attributes: UserID, CourseCategory, TimeSpentOnCourse, NumberOfVideosWatched, NumberOfQuizzesTaken, QuizScore, CompletionRate, and DeviceType. These attributes capture a range of information from user demographics and course-specific data to detailed engagement metrics.

The output attribute of interest is **CourseCompletionStatus**, which indicates whether a user completed the course (1) or did not complete it (0).

Data Sources:

The dataset was sourced from Kaggle. Here is the link:-



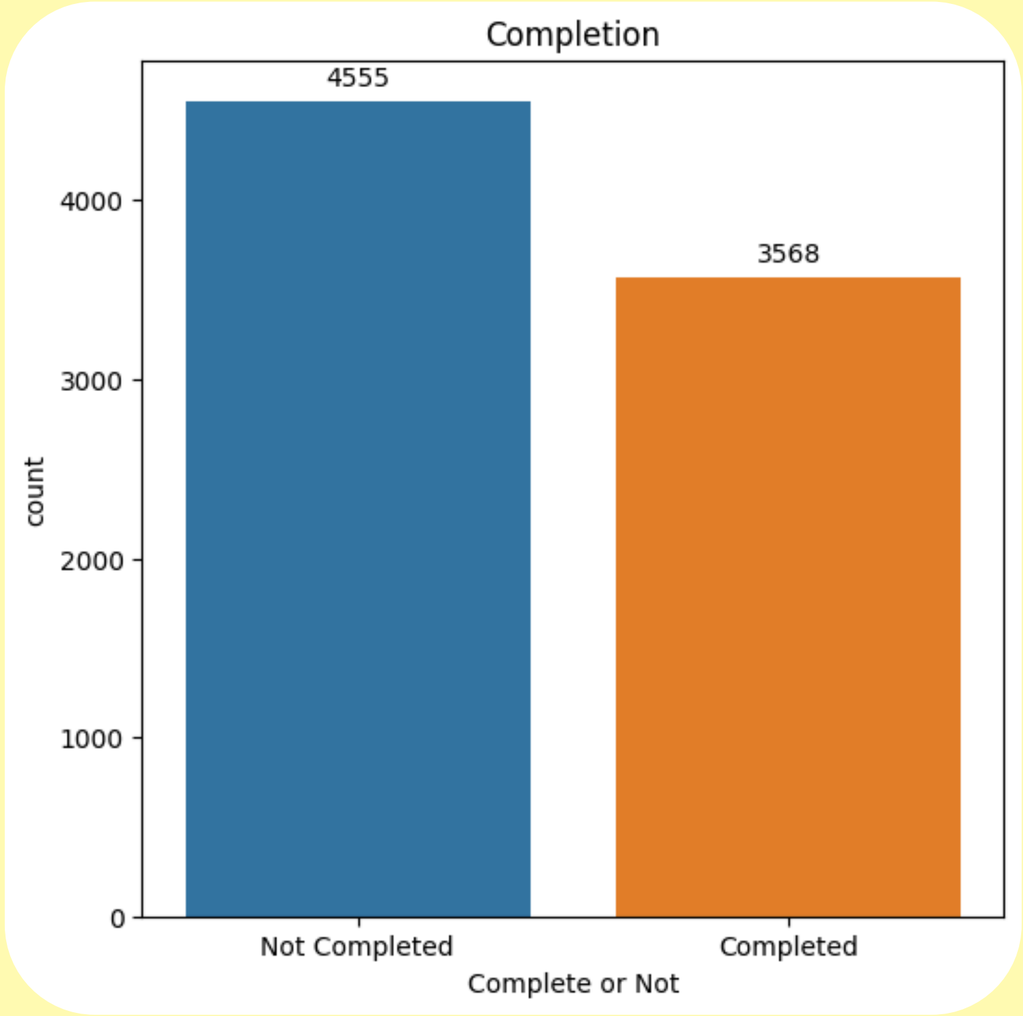
Data Collection and Cleaning

After loading the dataset, first step is to clean the data for making it suitable for further analysis. It includes:-

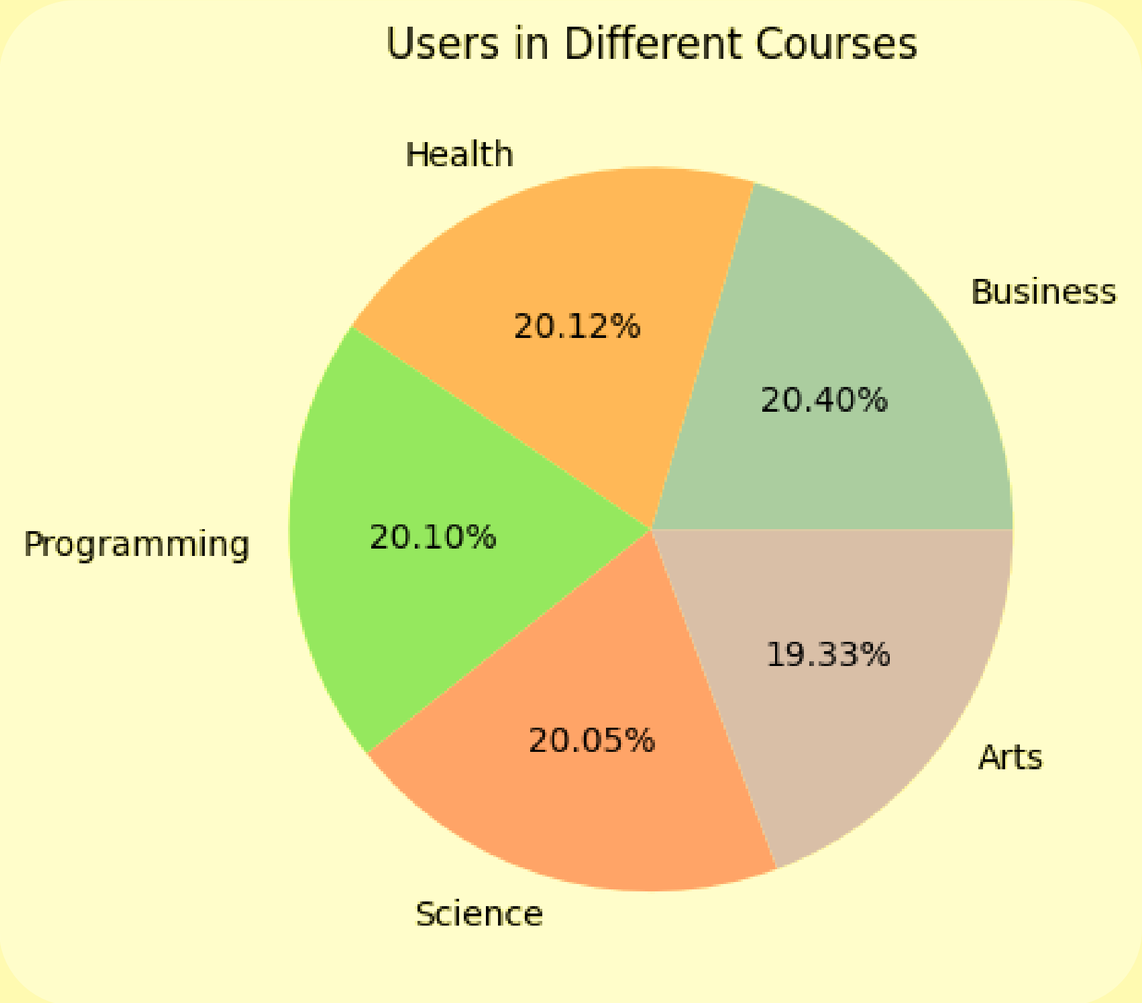
- **Identifying and Handling Missing Values**: This dataset does not contain any missing values and is already well-cleaned.
- **Duplicate Enteries:** The dataset contains 876 rows with duplicate entries, which have been removed.
- **Data Type Conversion:** Converting data types of columns to the appropriate format (e.g., converting date columns to datetime format).

Exploratory Data Analysis(EDA)

The number of users who have completed the course is less than the number of those who have not completed it.



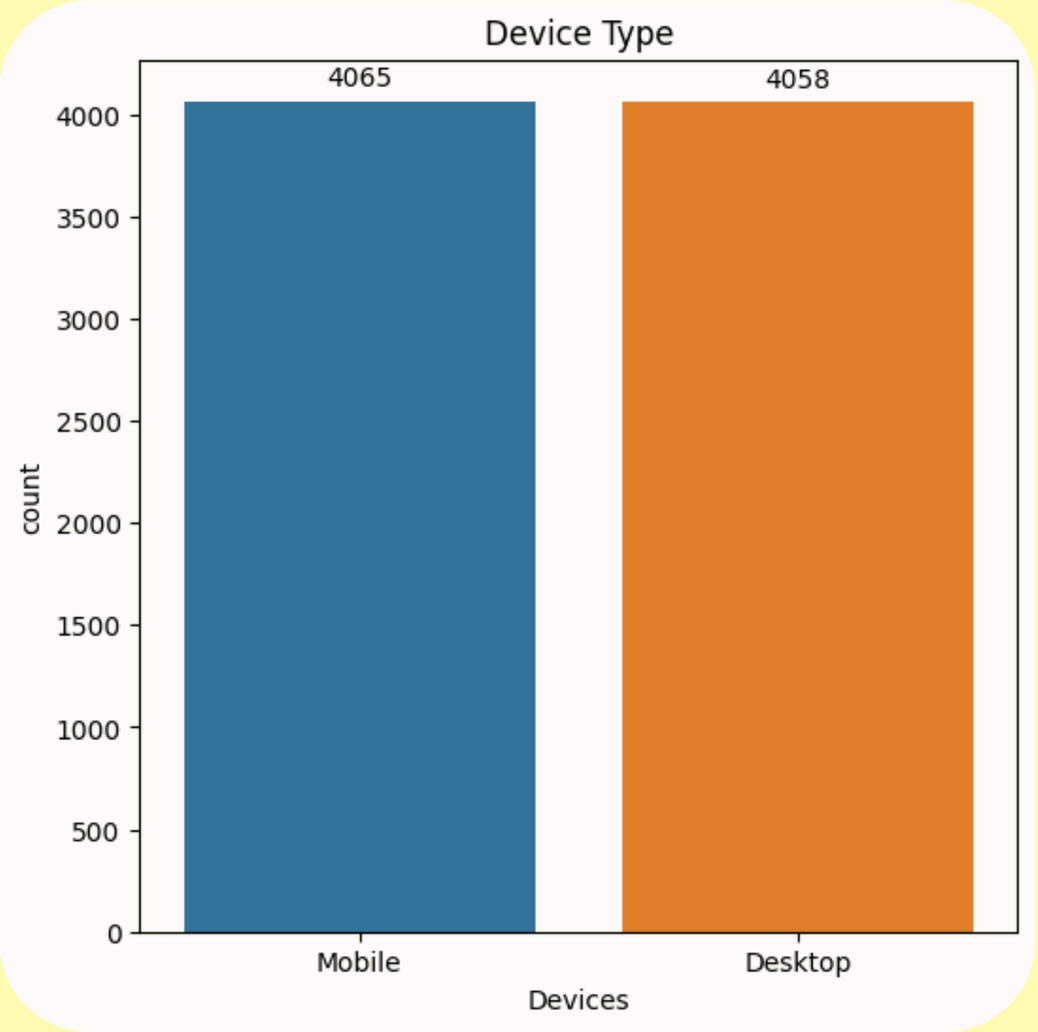
Exploratory Data Analysis(EDA)



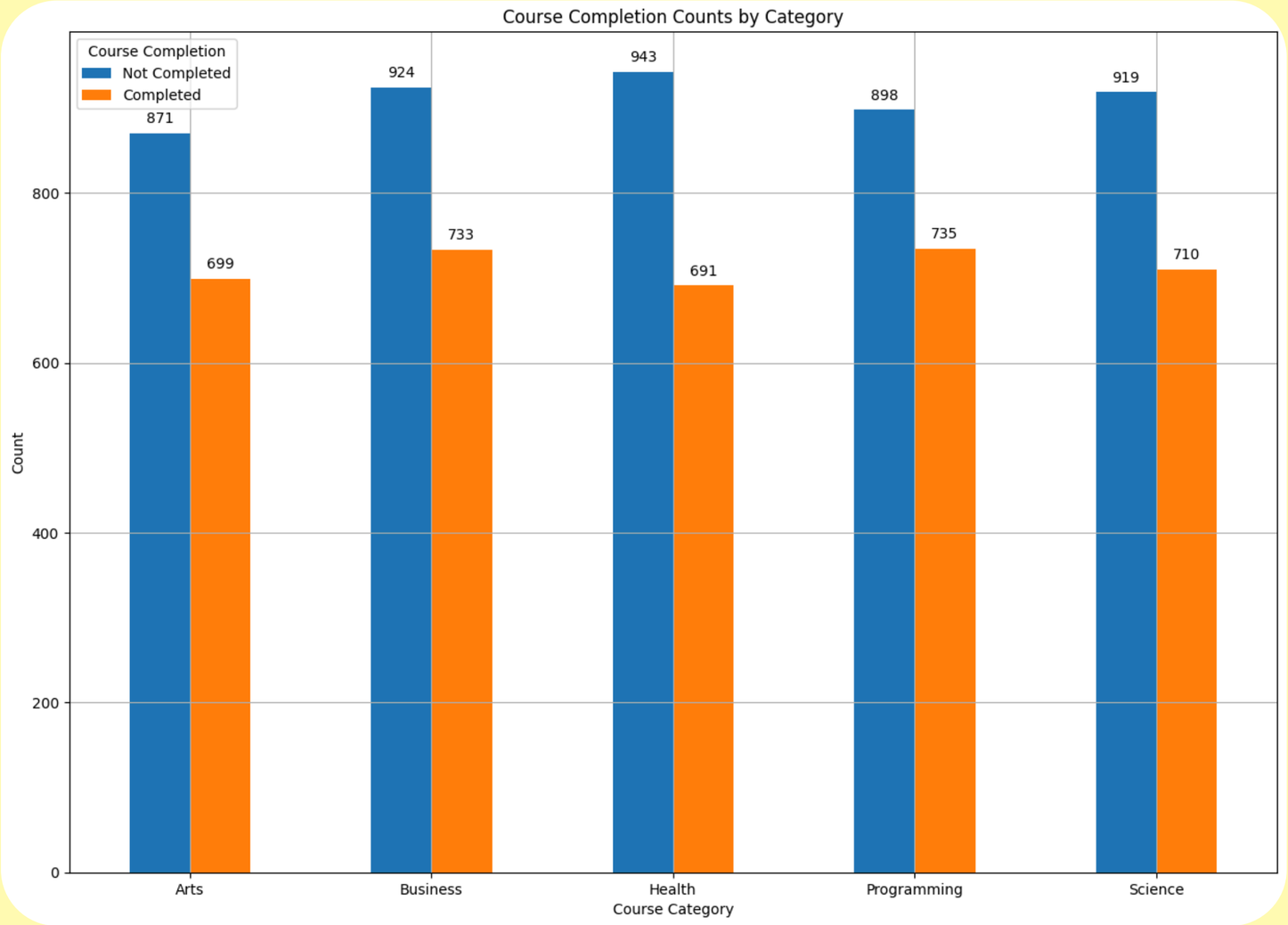
All users have an almost equal distribution of subjects, with the lowest representation in Arts and the highest in Business.

Exploratory Data Analysis(EDA)

There is little difference between the number of users using mobile phones and those using desktops.



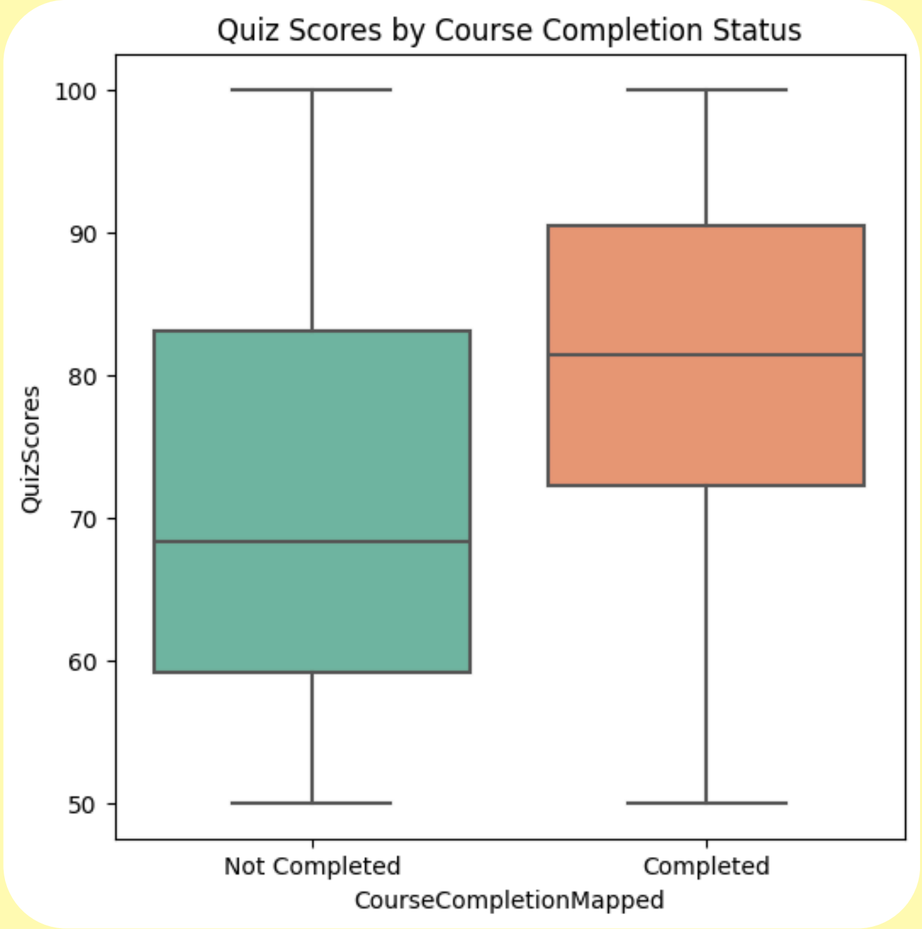
Exploratory Data Analysis(EDA)



No. of users who has completed the course are less than those who has not completed for all the categories of courses.

Exploratory Data Analysis(EDA)

The box plot shows that users who completed the course generally scored higher on quizzes compared to those who did not complete the course. This indicates that better quiz scores are often linked with successfully finishing the course.



Data Preparation

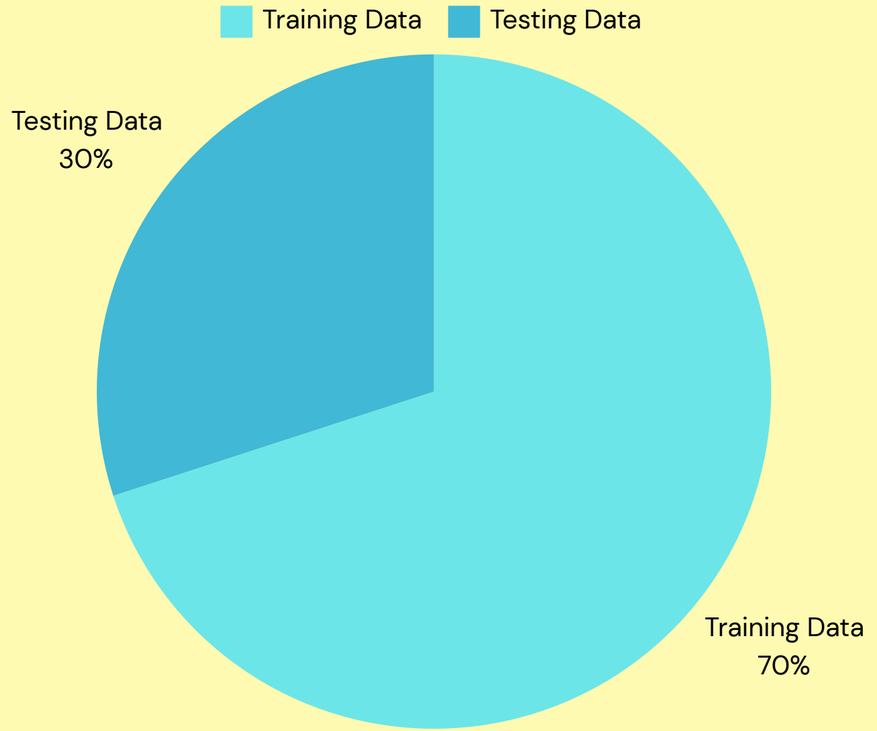
Selecting the relevant features and providing it to machine learning is import for accurate modelling. Here,

- The "**TimePerVideo**" feature was derived from TimeSpentOnCourse and NumberOfVideosWatched, but it contains many outliers, making it unsuitable for modeling and necessitating its removal..
- "The "**UserID**" feature was dropped because it is not relevant for predicting course completion.
- The "**CourseCategory**" variable was encoded into values 0, 1, 2, 3, and 4, corresponding to '**arts**', '**business**', '**health**', '**programming**', and '**science**', respectively.
- All continuous variables were **standardized** to ensure effective data preparation and improve modeling accuracy.

With these adjustments, the data is now ready for predictive analysis.

Training and Testing

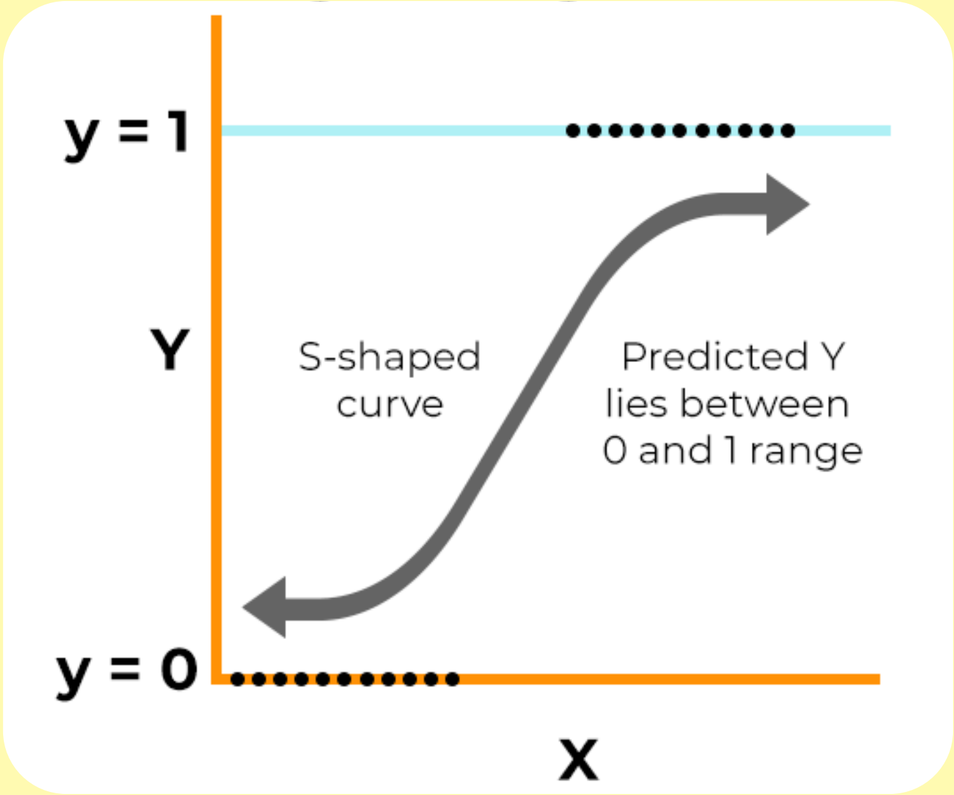
Splitting up the data into two parts training and testing set.
Training data refers to the set of examples used to train a machine learning model, which enables it to learn patterns and relationships within the data.
Testing data is use to assess the model's performance on unseen data.



Splitting up data

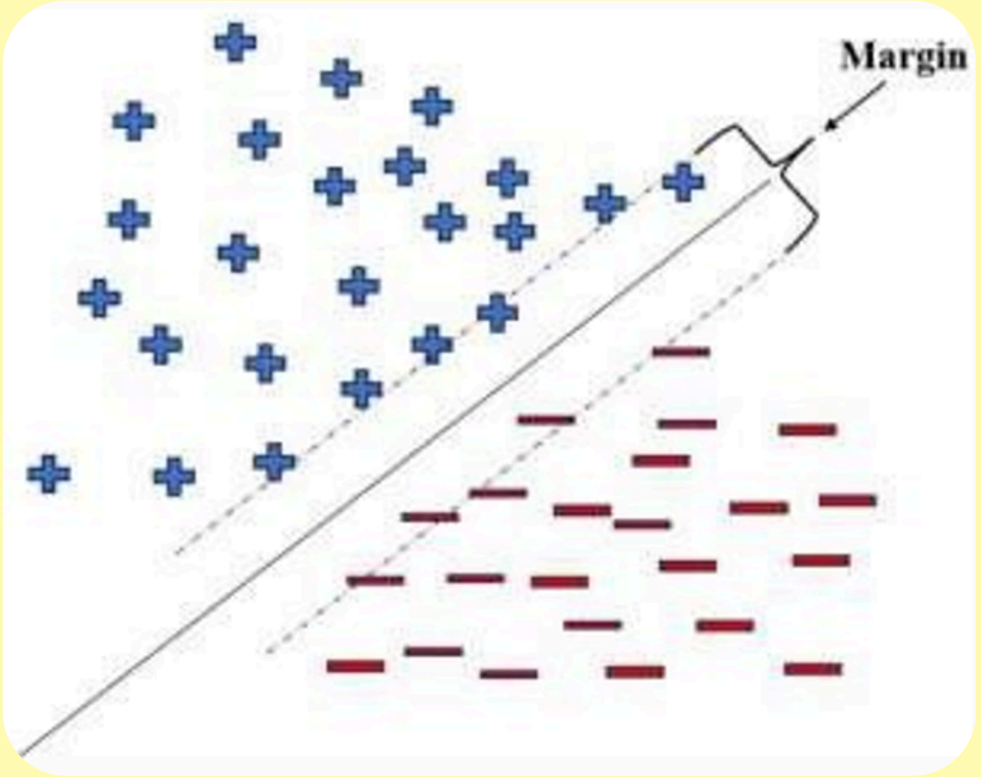
Machine Learning Models

Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.



Machine Learning Models

Support Vector Machine(SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.



Machine Learning Models

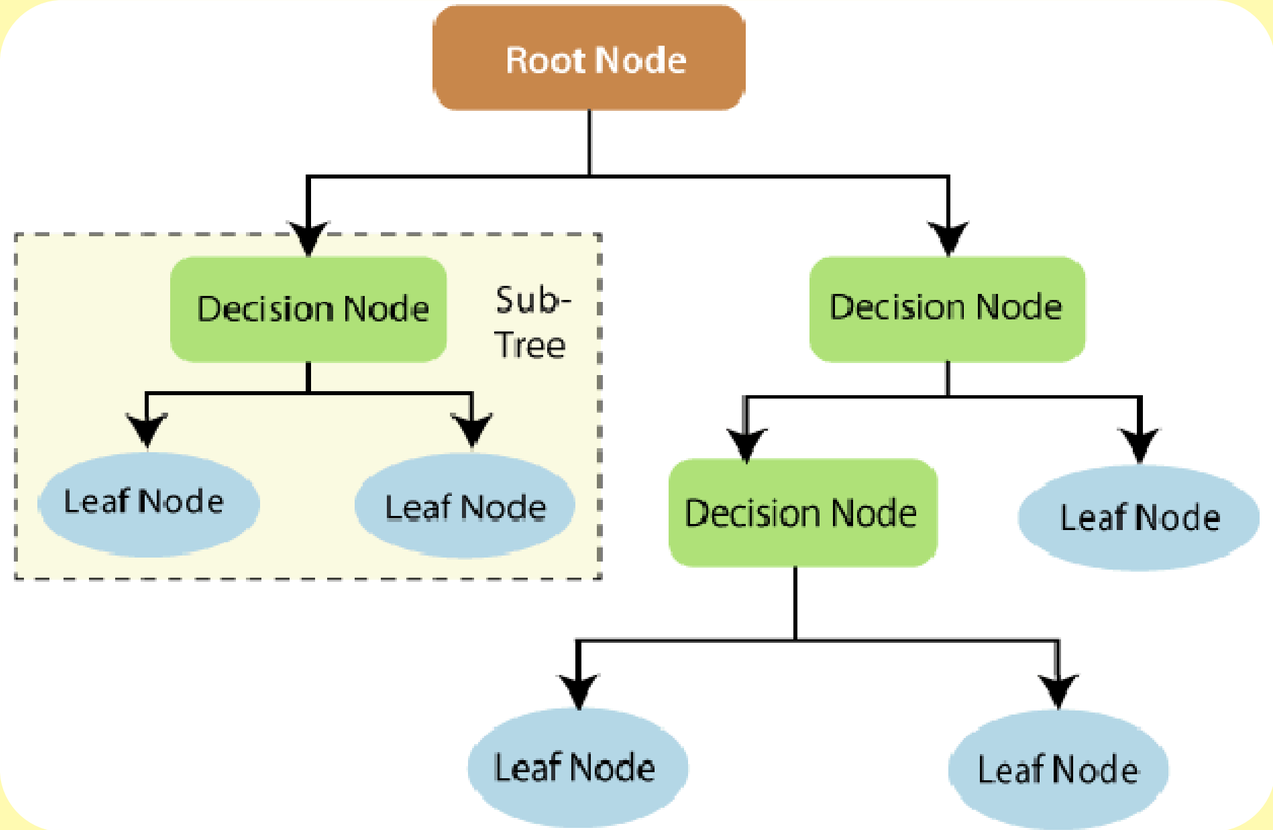
Naive Bayes uses Bayes' theorem to calculate the probability of a particular event based on prior knowledge of conditions that might be related to that event. In the context of classification, it calculates the probability of a given class label (C) based on observed feature values (X).

$$P(C|X) = \frac{P(C).P(X|C)}{P(X)}$$

For continuous data, **Gaussian Naive Bayes** applies this principle by assuming that the feature values follow a Gaussian (normal) distribution. This allows it to estimate the probability of the class label by considering the distribution of the continuous features.

Machine Learning Models

Decision tree classifier is a supervised learning algorithm that uses a tree-like model to classify data. It works by recursively partitioning the data into subsets based on the most informative features. Each internal node represents a feature or attribute, while each leaf node represents a class label. The algorithm splits the data into subsets until all examples in a subset belong to the same class.



Evaluation Parameters

Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model**.**

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is a metric that measures how often a machine learning model correctly identifies positive instances (true positives) from all the actual positive samples in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

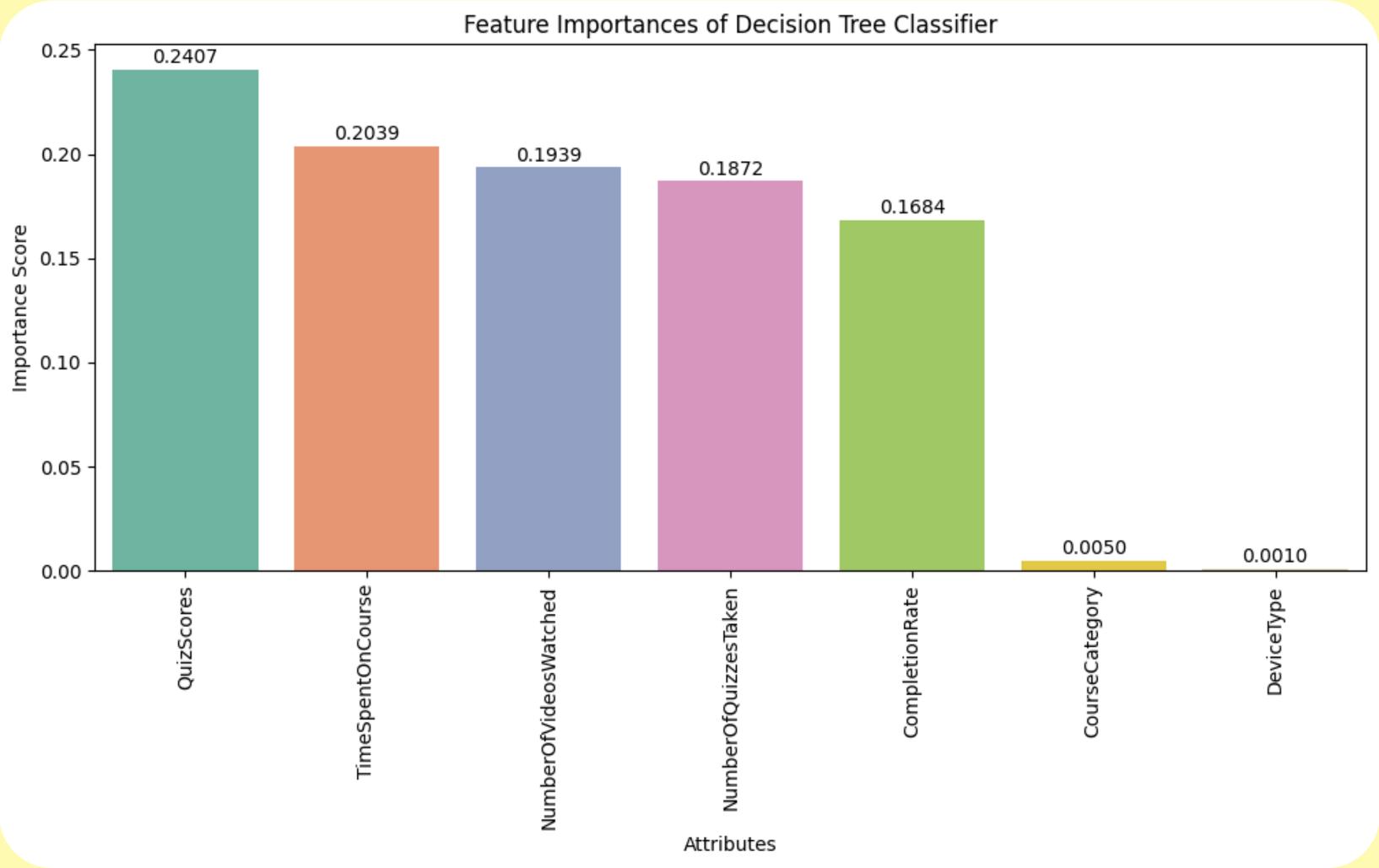
Why these measures?

- **Training accuracy** is used to verify that our models are not **overfitted**. Overfitting occurs when a model performs well on training data but poorly on test data, so testing accuracy helps ensure that the model generalizes well to new data.
- **Understanding and minimizing false positives helps to identify why some of the users do not finish the courses**. This insight allows for targeted guidance and motivation to improve their chances of successful course completion which makes the **precision** as important measure to be consider.
- **Recall** values helps in **identifying the users who has actually completed the course** providing them more opportunities to support and recognize their achivements, potentially impacting their motivation and satisfaction.
- Since we consider both precision and recall, the **F1 score is crucial for evaluating the trade-off between these metrics**, providing a balanced measure of model performance.

Experimental Results

	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Logistic Regresion	0.7932	0.7887	0.7693	0.7472	0.7581
Support Vector Machine(SVM)	0.9286	0.8913	0.8906	0.8601	0.8751
Decision Trees	0.9601	0.9520	0.9607	0.9296	0.9448
Gaussian Naive Bayes	0.8241	0.8252	0.8378	0.7509	0.7920

Feature Importance of the Best Classifier



Conclusion

- **Decision Trees** and **SVM** demonstrate the most robust and balanced performance, making them the preferred models for predicting course completion. They achieve high accuracy, precision, recall, and F1-scores, ensuring effective identification of students likely to complete the course while minimizing incorrect predictions.

- The feature importance plot of the Decision Tree classifier suggests **a well-balanced model where no single feature dominates,** with multiple features contributing meaningfully to predictions**.**