

On

“Predict Online Learning Completion”

**BACHELOR OF TECHNOLOGY**

**DEGREE**

SESSION 2024-25

in

**CSE(AIML)**

By

Name : Preeti Singh

Roll Number : 202401100400146

Section: B

**Under the supervision of**

“ABHISHEK SHUKLA”

**KIET Group of Institutions, Ghaziabad**

**May, 2025**

## **1. Introduction**

As online education platforms become increasingly popular, understanding student engagement and success is essential. This project aims to automate the prediction of course completion using supervised machine learning. By leveraging a dataset containing student activity data—such as videos watched, assignments submitted, and forum interactions—we build a model to help educators identify at-risk students early and take proactive steps.

## **2. Problem Statement**

To predict whether a student will complete an online course based on their activity during the course. Accurate predictions can help course providers enhance student support, personalize learning experiences, and improve overall retention rates.

## **3. Objectives**

* Preprocess the dataset for training a machine learning model.
* Train a **Random Forest Classifier** to classify student course completion.
* Evaluate model performance using standard classification metrics.
* Visualize the confusion matrix using a heatmap for better interpretability.

## **4. Methodology**

**Data Collection:** The dataset online\_learning.csv was uploaded, containing information about student interactions with course content.

**Data Preprocessing:**

* Convert the target column completed from 'yes'/'no' to binary (1/0).
* No missing values required imputation.
* Since all features are numerical, no one-hot encoding was needed.
* Scale feature values using StandardScaler to normalize inputs.

**Model Building:**

* Split the dataset into 80% training and 20% testing sets.
* Train a **Random Forest Classifier**, a robust ensemble learning model for classification tasks.

**Model Evaluation:**

* Assess the model using accuracy, precision, recall, and confusion matrix.
* Visualize the confusion matrix using Seaborn’s heatmap for easier interpretation of performance.

## **5. Data Preprocessing**

The following preprocessing steps were applied:

* Verified that there were no missing values to handle.
* All input features (videos\_watched, assignments\_submitted, forum\_posts) were numerical.
* Applied StandardScaler to standardize the features.
* Used train\_test\_split to split the dataset into training (80%) and testing (20%) sets.

## **6. Model Implementation**

A **Random Forest Classifier** was chosen for this task due to its effectiveness in handling classification problems with tabular data. The model was trained using the scaled training dataset and then used to predict the test set outcomes.

## **7. Evaluation Metrics**

The following metrics were used to evaluate model performance:

* **Accuracy**: Measures the proportion of correct predictions.
* **Precision**: Indicates how many of the predicted completions were actual completions.
* **Recall**: Measures how many actual completions were correctly predicted.
* **F1 Score** (optional for deeper analysis): Harmonic mean of precision and recall.
* **Confusion Matrix**: A visualization tool to show the distribution of predictions and actual outcomes.

## **8. Results and Analysis**

* The trained model achieved high accuracy and balanced precision and recall on the test data.
* The confusion matrix revealed how well the model differentiated between students who completed and did not complete the course.
* Visual inspection of the heatmap provided an intuitive understanding of correct versus incorrect prediction

## **9. Conclusion**

The Random Forest model effectively classified whether students would complete an online course based on activity features. This predictive capability can help online education platforms proactively support learners and improve course completion rates. Future work can explore more complex models (e.g., Gradient Boosting, XGBoost) or address potential class imbalance issues if present.

## **10. References**

* scikit-learn documentation:<https://scikit-learn.org>
* pandas documentation: https://pandas.pydata.org
* Seaborn visualization library: https://seaborn.pydata.org
* Research literature on educational data mining and student retention





