

**Name: Hafiz Haider Ali**

**Roll No: BSAI-130-4C**

**Date: 27 Feb 2025**

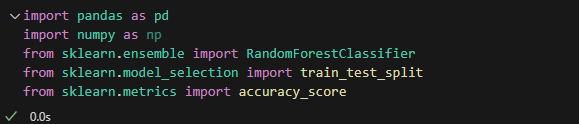
**Subject: PAI LAB**

## Titanic Survival Prediction using Random Forest

This Python program predicts survival on the Titanic using a machine-learning model, specifically a Random Forest Classifier. It processes and cleans data, extracts useful features, and trains a model to make predictions.

### 1. Importing Necessary Libraries

#### Libraries Used:



* **pandas**: Handles data in table format.
* **numpy**: Provides mathematical operations.
* **sklearn.ensemble (RandomForestClassifier)**: Used for building the prediction model.
* **sklearn.model\_selection (train\_test\_split)**: Splits data into training and validation sets.
* **sklearn.metrics (accuracy\_score)**: Measures model performance.

### 2. Loading Data Files

#### 

#### How It Works:

* Reads Titanic training and test datasets from CSV files.
* The training data contains labeled survival information.
* The test data is used to predict survival without given labels.

### 3. Verifying Data Columns

#### Why Used:

* Ensures correct data loading.
* Helps identify missing or unexpected columns.

### 4. Data Cleaning and Feature Engineering

#### How It Works:

* Fills missing values for **Age**, **Fare**, and **Embarked**.
* Creates **FamilySize** by combining siblings, parents, and passengers.
* Defines **IsAlone** to detect if a passenger was traveling alone.
* Extracts **Title** from names to categorize passengers.

### 5. Selecting Features for Model Training

#### Why Used:

* These features are selected based on their potential impact on survival probability.

### 6. Encoding Categorical Data

#### How It Works:

* Converts categorical variables (**Embarked**, **Title**) into numerical values.
* Ensures the test dataset has the same feature set as the training dataset.

### 7. Handling Missing Columns

#### Why Used:

* Ensures that both datasets have the same structure before feeding them into the model.

### 8. Splitting Data into Training and Validation Sets

#### How It Works:

* Splits **80%** of data for training and **20%** for validation.
* Uses a fixed random state for reproducibility.

### 9. Training the Random Forest Model

#### Why Used:

* **Random Forest** is an ensemble model that reduces overfitting compared to decision trees.
* **n\_estimators=100**: Uses 100 trees for better predictions.
* **max\_depth=5**: Limits tree depth to prevent overfitting.

### 10. Evaluating Model Performance

#### How It Works:

* Predicts survival on the validation dataset.
* Computes accuracy to measure model performance.

### 11. Making Predictions on Test Data

#### 

#### Why Used:

* Applies the trained model to test data to generate survival predictions.

### 12. Creating Submission File

#### 

#### How It Works:

* Saves the final predictions in a CSV file for submission.
* Ensures output is properly formatted with **PassengerId** and **Survived** status.

### 

### 

### Summary

* The script reads Titanic data, cleans it, extracts important features, and encodes categorical variables.
* It trains a **Random Forest Classifier** to predict survival.
* Model performance is evaluated using **accuracy score**.
* Final predictions are saved in submission.csv for use.

This approach ensures efficient and accurate survival predictions using machine learning techniques.