

**Superior University Gold Campus**

**PAI Lab Task #2**

Name: Syed Ejiz Ul Hassan Kazmi

Roll. No: SU92-BSDSM-F23-025

Section: BSDS-4A

Instructor: Sir Rasikh

**House Prediction Model**

## **1. Data Loading:**

 The dataset is loaded using **Pandas** with pd.read\_csv().

 The dataset is stored in a variable named df.

 df.info() is used to check the structure, including column names and missing values.

## **2. Data Exploration:**

Various checks are performed to understand the dataset:

* data.isnull().sum(): Counts missing values in each column.
* data.columns: Lists all column names.
* len(data.columns): Prints the number of columns.
* data.duplicated().sum(): Counts duplicate rows.

## **3. Data Preprocessing:**

Data preprocessing involves cleaning and transforming raw data for machine learning models.

### ****3.1 Dropping Unnecessary Columns:****

Some columns are dropped to reduce noise and improve model efficiency.

'PassengerId', 'Name', and 'Cabin' are removed as they do not contribute to predictive analysis.

**3.2 Handling Missing Values:**

Different imputation techniques are applied to fill missing values:

* **Numerical columns**: Missing values are replaced with the **median**.
* **Categorical columns**: Missing values are replaced with the **mode (most frequent value)**.

**3.3 Feature Encoding:**

* Categorical variables are converted into numerical form using **label encoding**.

Label encoding converts text-based categories into numbers for machine learning models.

## **4. Machine Learning Model Used:**

The notebook trains a **Random Forest Classifier** to predict an outcome based on the features.

**4.1 Splitting Data**

* The dataset is divided into **features (X)** and **target (y)**.
* train\_test\_split() is used to split the dataset into **training (80%)** and **testing (20%)** sets.

**4.2 Training the Model**

A **Random Forest Classifier** is trained using the processed dataset.

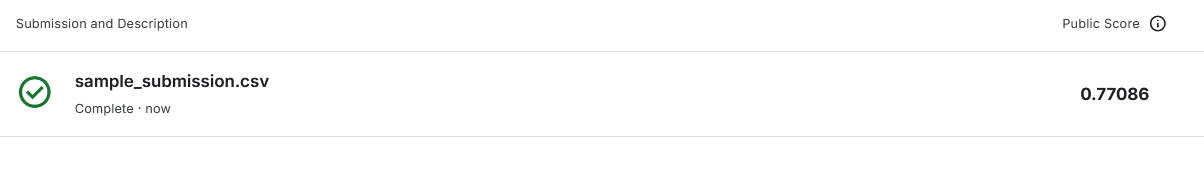
 n\_estimators=100: The model uses 100 decision trees for classification.

 The model is trained on X\_train and y\_train.

**4.3 Evaluating Model Performance**

* The model is tested on the test set, and accuracy is calculated.

**5.Score on kaggle:**

****

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**