

**Superior University Gold Campus**

**PAI Lab Task #1**

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 the Pandas library with the pd.read\_csv() function.
* It reads data from a CSV file into a Pandas DataFrame, allowing further analysis.

## **2. Data Exploration:**

The following techniques are used to understand the dataset:

* df.describe(): Provides summary statistics for numerical columns.
* df.info(): Displays data types, column names, and non-null values.
* df.isnull().sum(): Counts the number of missing values in each column.
* df.columns: Lists all column names in the dataset.
* cor\_mat = df.corr(): Computes the correlation matrix to identify relationships between variables.
* sales\_price = cor\_mat['SalePrice']: Extracts the correlation of all features with the target variable **SalePrice**.

## **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.

**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 Data Type Conversion****

Categorical variables are converted into numerical format for machine learning algorithms.

### ****3.5 Splitting Data****

The dataset is split into **training** and **test** sets to train and evaluate the model.

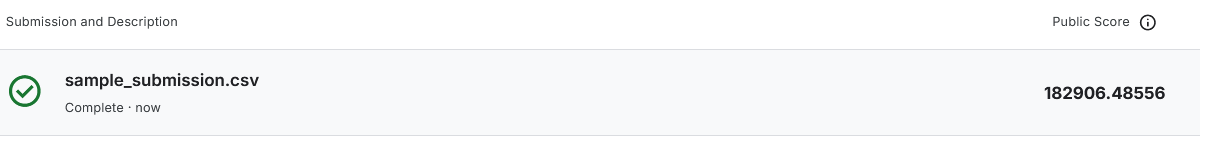
## **4. Machine Learning Model Used:**

The code applies **Random Forest Regressor**, an ensemble learning method for regression

**Why Random Forest?**

* It builds multiple decision trees and averages their results to improve accuracy.
* It reduces overfitting by using different subsets of data and features.
* It works well with large datasets and complex relationships.

**5.Score on kaggle:**

****

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