****

**PROGRAMMING FOR AI (LAB)**

**Lab Task No 01**

**Submitted To :- Sir Rasikh Ali**

**Submitted By :- Aiman Ijaz**

**Roll No :- SU92-BSAIM-F23-010**

**Department :- Software Engineering**

**Section :- BSAI-4A**

**Housing Prices Competition for Kaggle Learn Users**

**Project Documentation**

**1. Introduction**

This project aims to predict house prices using a machine learning model. The dataset consists of various features related to house characteristics, such as size, location, number of rooms, and other structural attributes. The model is trained using a **Random Forest Regressor**, which is an ensemble learning technique.

**2. Data Exploration**

**2.1 Dataset Overview**

Two datasets are used in this project:

**Train Dataset (train.csv)**: Contains labeled data with house prices.

**Test Dataset (test.csv)**: Contains unlabeled data for making predictions.

**2.2 Data Analysis**

The following operations are performed to explore the dataset:

Display the first and last few rows of the data using .head() and .tail().

Obtain summary statistics using .describe().

Check for missing values using .isnull().sum().

Identify unique values in categorical features using .unique().

Check the number of instances using .count().

**3. Data Preprocessing**

Before training the model, data preprocessing is necessary to handle missing values and convert categorical data into numerical format.

**3.1 Handling Missing Values**

**Categorical features**: Missing values are filled using the **mode** (most frequent value).

**Integer features**: Missing values are replaced with the **mode**.

**Float features**: Missing values are replaced with the **mean** of the respective column.

**3.2 Data Type Conversion**

Some float features are converted into integers for better compatibility using .astype('int64').

**3.3 Encoding Categorical Data**

Categorical columns are converted into numerical values using **Label Encoding**, where each unique category is assigned a numerical label.

**4. Model Training**

**4.1 Splitting the Data**

The target variable **SalePrice** is separated from the training data.

Features from the test data are aligned with the training dataset columns.

**4.2 Training the Model**

A **Random Forest Regressor** is used to train the model.

The number of estimators is set to **100**, and a random seed (random\_state=42) ensures reproducibility.

The model learns patterns from the training data.

**5. Making Predictions**

The trained model predicts house prices for the test dataset.

The predictions are stored in a **submission file (submission.csv)** containing the house ID and predicted price.

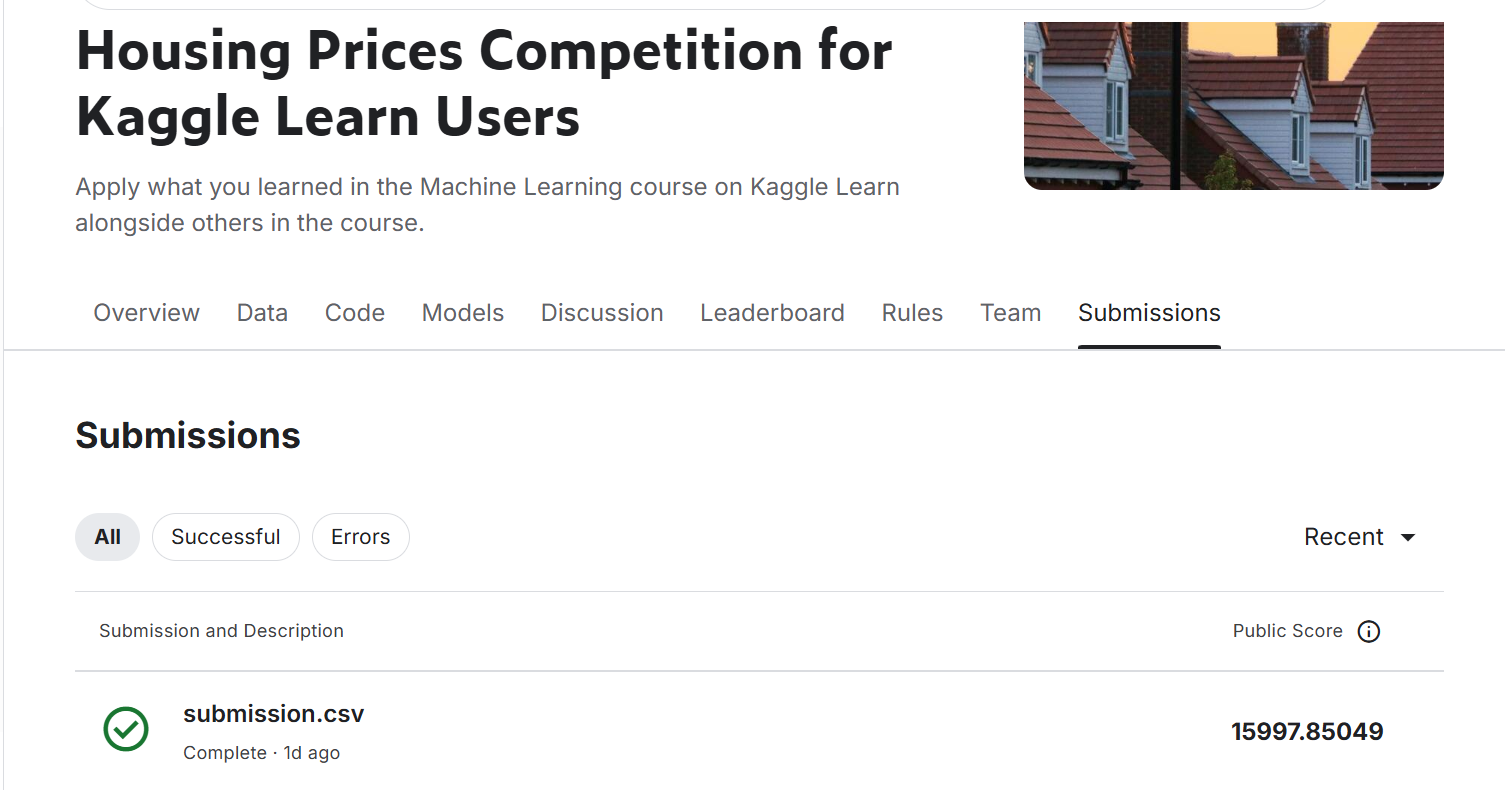
**6. Conclusion**

This project successfully builds a machine learning model to predict house prices based on various housing attributes.

Data preprocessing, including handling missing values and encoding categorical data, plays a crucial role in improving model performance.

The **Random Forest Regressor** is an effective algorithm for this task as it reduces overfitting and provides robust predictions.

**Output:**

****