California housing price prediction

# Introduction

Regression involves using one or more variables, labelled independent variables, to predict the values of another variable, the dependent variable. The purpose of the project is to predict median house values in Californian districts, given many features from these districts.

**Main Objective**

* The prime objective of this project is to construct a working model which has the capability of predicting the value of houses.

**Domain:** Finance and Real estate

**Data Description**

* Description: The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

**Data Dictionary – Variable and Description**

● longitude (signed numeric - float) : Longitude value for the block in California, USA

● latitude (numeric - float ) : Latitude value for the block in California, USA

● housing\_median\_age (numeric - int ) : Median age of the house in the block

● total\_rooms (numeric - int ) : Count of the total number of rooms (excluding bedrooms) in all houses in the block

● total\_bedrooms (numeric - float ) : Count of the total number of bedrooms in all houses in the block

● population (numeric - int ) : Count of the total number of population in the block

● households (numeric - int ) : Count of the total number of households in the block

● median\_income (numeric - float ) : Median of the total household income of all the houses in the block

● ocean\_proximity (numeric - categorical ) : Type of the landscape of the block [ Unique Values : 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND' **]**

● **median\_house\_value** (numeric - int ) **:** Median of the household prices of all the houses in

the block

**Dataset Size:** 20640 rows x 10 columns

**Project Guidelines :**

1. **Load the data :**

● Read the **“housing.csv”** file from the folder into the program.

● Print first few rows of this data.

● Extract input (X) and output (Y) data from the dataset.

2. **Handle missing values :**

● Fill the missing values with the median of the respective column.

3. **Encode categorical data :**

● Convert categorical column in the dataset to numerical data.

4. **Split the dataset :**

● Split the data into 80% training dataset and 20% test dataset.

5. **Standardize data :**

● Standardize training and test datasets.

6. **Perform Linear Regression :**

● Perform Linear Regression on training data.

● Predict output for test dataset using the fitted model.

● Print root mean squared error (RMSE) from Linear Regression.

[ HINT: Import **mean\_squared\_error** from **sklearn.metrics** ]

7. **Perform Decision Tree Regression :**

● Perform Decision Tree Regression on training data.

● Predict output for test dataset using the fitted model.

● Print root mean squared error from Decision Tree Regression.

8. **Perform Random Forest Regression :**

● Perform Random Forest Regression on training data.

● Predict output for test dataset using the fitted model.

● Print RMSE (root mean squared error) from Random Forest Regression.

**9.Saving the model and pipeline**

* Saving both the model and the final pipeline to pickle file for future use