Predicting House Prices using Machine Learning

# INTRODUCTION

# Accurately estimating the value of real estate is an important problem for many stakeholders including house owners, house buyers, agents, creditors, and investors. It is also a difficult one. Though it is common knowledge that factors such as the size, number of rooms and location affect the price, there are many other things at play. Additionally, prices are sensitive to changes in market demand and the peculiarities of each situation, such as when a property needs to be urgently sold. The sales price of a property can be predicted in various ways, but is often based on regression techniques. All regression techniques essentially involve one or more predictor variables as input and a single target variable as output

# DATASET

The data is obtained from https://www.kaggle.com/datasets/vedavyasv/usa-housing

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Avg. Area Income | 1.000000 | -0.002007 | -0.011032 | 0.019788 | -0.016234 | 0.639734 |
| Avg. Area House Age | -0.002007 | 1.000000 | -0.009428 | 0.006149 | -0.018743 | 0.452543 |
| Avg. Area Number of Rooms | -0.011032 | -0.009428 | 1.000000 | 0.462695 | 0.002040 | 0.335664 |
| Avg. Area Number of Bedrooms | 0.019788 | 0.006149 | 0.462695 | 1.000000 | -0.022168 | 0.171071 |
| Area Population | -0.016234 | -0.018743 | 0.002040 | -0.022168 | 1.000000 | 0.408556 |
| Price | 0.639734 | 0.452543 | 0.335664 | 0.171071 | 0.408556 | 1.000000 |

# COLUMNS USED

##### 

##### From USA\_Housing.csv data the following columns are used

* **Avg. Area Income**
* **Avg. Area House Age**
* **Avg. Area Number of Rooms**
* **Avg. Area Number of Bedrooms**
* **Area Population**
* **Price**
* **Address**

# LIBRARIES USED

The Python 3 environment comes with many helpful analytics libraries installed and several helpful packages to load.

The essential libraries used in this project are :

* Importing OS (for kaggle inputs)
* Numpy and Pandas libraries
* Matplotlib
* Seaborn

# TRAIN AND TEST

Training the dataset by describe(), isnull().sum(), drop(), show(), and by using k-means algorithm we train the data

Testing the data by importing sklearn.cluster from k-means with ensuring the plot range and axis labels producing the k value, scattering the data by kmeans.cluster\_centers and producing 3D

plot.

# REST OF THE EXPLANATIONS

Data Collection

The process involves gathering house prices, which includes information about their purchase history, demographics, and interaction patterns.

Data Preprocessing

The task involves preparing and cleaning data, handling missing values, and converting categorical features into numerical representations.

Feature Engineering

Data preparation and cleaning, handling missing values, and the transformation of categorical features into numerical representations are all part of the task.

# ALGORITHMS USED

Apply clustering algorithms like K-Means, DBSCAN, or hierarchical clustering to segment customers.

Visualization: Visualize the customer segments using techniques like scatter plots, bar charts, and heatmaps. Interpretation: Analyze and interpret the characteristics of house prices to derive actionable insights for marketing strategies.

# DESIGN AND DATA FLOW CHART:



