**HOUSE PRICE PREDICTION**

**PROBLEM DEFINITION:**

The problem is to predict house prices using machine learning techniques. The objective is to develop a model that accurately predicts the prices of houses based on a set of features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors. This project involves data preprocessing, feature engineering, model selection, training, and evaluation.

**DESIGN THINKING:**

**1.Data Source:**

We choose a dataset from Kaggle which contains information about house including features like location, square footage, number of bedrooms, average number of rooms, and price.

**Avg. Area Income**: This column represents the average income of residents in the area where the property is located.

**Avg. Area House Age**: This column represents the average age of houses in the same area.

**Avg. Area Number of Rooms**: This column indicates the average number of rooms in houses in the area.

**Avg. Area Number of Bedrooms:** This column represents the average number of bedrooms in houses in the area.

**Area Population:** This column represents the population of the area where the property is situated.

**Price:** This column contains the price of the property.

**Address:** This column provides the address of each property, including city or town names, postal codes, and unit or apartment numbers.

**Dataset Link:**[**https://www.kaggle.com/datasets/vedavyasv/usa-housing**](https://www.kaggle.com/datasets/vedavyasv/usa-housing)

**2. Data Preprocessing:**

Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.

**Check for missing values**: *isnull()* function to identify missing values in the dataset and *sum()* to count them for each column.

**Handle missing values:** Fill missing values with the mean of the respective columns using the *fillna()* method.

**Convert categorical features into numerical representations:** One hot Encoding is the best way to convert categorical data into binary vectors. This maps the values to integer values.

**3.Feature Selection:**

For predicting house prices **feature variable** as X which contains **'Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',' Avg. Area Number of Bedrooms', 'Area Population'** and target variable as Y which contains **‘Price’.**

**4.** **Model Selection:**

 A suitable regression algorithm for predicting house prices is **Linear Regression**. Linear Regression predicts the final output-dependent value based on the given independent features. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable.

**5. Model Training:**

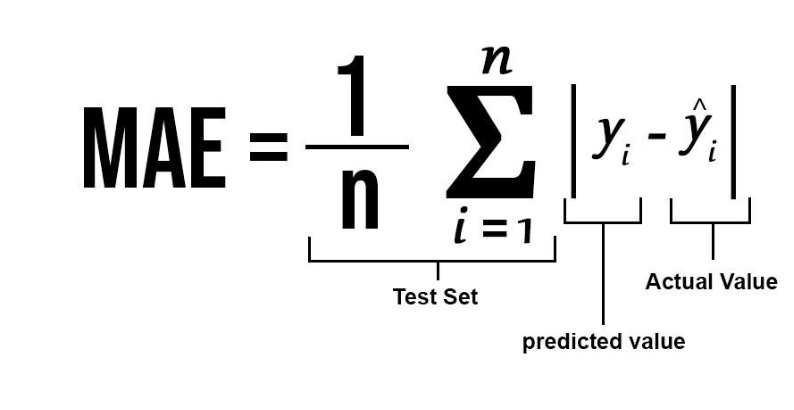
Train the selected model using the preprocessed data. Feed the training data (features and labels) into the model. The model will adjust its internal parameters to learn the underlying patterns in the data. During training, the algorithm chosen tries to minimize a certain loss function, which measures the difference between its predictions and the true labels in the training data.

**6. Evaluation:**

To evaluate the model's performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

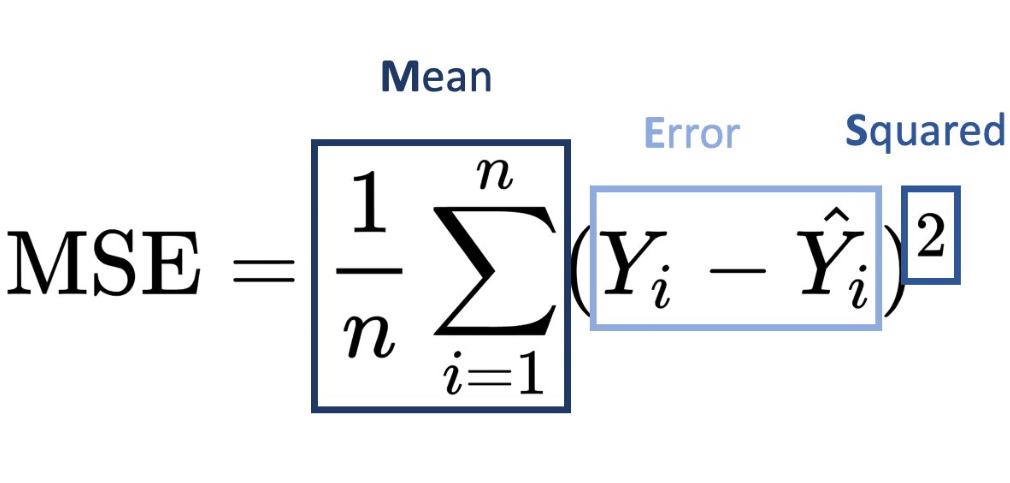
**MAE:**

Mean Absolute Error (MAE) is a straightforward metric for assessing the accuracy of a regression model. It provides a clear and interpretable measure of how well the model predicts continuous numerical value.

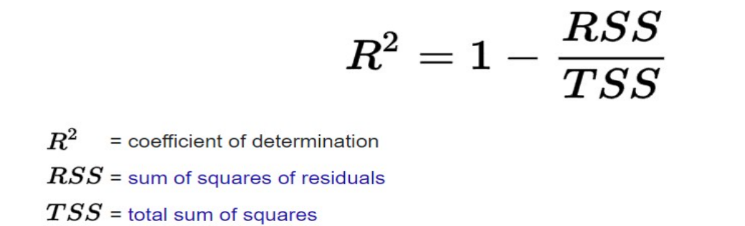


**MSE:**

MSE stands for "Mean Squared Error." It is another common metric used for evaluating the performance of regression models, especially in the context of predictive modeling. MSE quantifies the average of the squared differences between predicted values and actual (true) values in a dataset.



**R-SQUARED:**

R-squared (R²), also known as the coefficient of determination, is a statistical measure used to assess the goodness of fit of a regression model. It quantifies the proportion of the variance in the dependent variable (target variable) that is explained by the independent variables (features) in the model. In other words, R-squared indicates how well the model fit the data.