**House Price Prediction**

**Problem Statement:**

The problem statement is to predict house prices using machine learning techniques. Our goal is to create a model that can accurately estimate the prices of houses based on a set of key features, including location, square footage, number of bedrooms, number of bathrooms, and other relevant factors.

**Design Thinking:**

**1. Data Source Selection:**

Choose a reliable dataset containing house features like location, square footage, bedrooms, bathrooms and corresponding prices from real estate databases or public datasets.

**2. Data Preprocessing:**

**Data Cleaning**:

Handle missing data, remove outliers, and correct inaccuracies.

**Data Transformation:**

Convert categorical features to numerical representations using encoding techniques.

**3. Feature Selection:**

**Feature Correlation Analysis**:

Identify features strongly correlated with house prices.

**Feature Importance Scores**:

Use models to determine the most influential features.

**4. Model Selection:**

The choice of a regression model is important for accurate price prediction

**Linear Regression:** Simple and interpretable.

**Random Forest Regressor:** Suitable for capturing complex relationships.

Choose the model based on dataset size, complexity, and interpretability requirements.

**5. Model Training:**

Data Split: Divide the dataset into training and testing sets.

Model Fitting: Train the chosen model, including parameter optimization for performance.

**6. Evaluation:**

To determine the effectiveness of our predictive model, we will employ relevant regression metrics such as:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R-squared (R^2)

**Phases of Development**

**Phase 1:**

**Objective:** Utilize machine learning techniques to predict house prices.

**Design Thinking Approach:**

* Data Source Selection
* Data Preprocessing
* Feature Selection
* Model Selection
* Model Training
* Evaluation

**Phase 2:**

**Objective:** Improve the existing model using deep learning techniques.

**Key Actions:**

Explore the dataset to understand its structure and preprocess it.

Design a neural network with specific layers and activation functions.

Train the model and evaluate its performance on the test dataset.

**Phase 3:**

**Objective:** Loading the dataset and preprocessing for analysis.

**Key Steps:**

Implementing the Design by coding the predictive model.

* Import the dataset
* Import necessary libraries.
* Preprocessing the data: handling missing values, feature selection, and scaling.

**Phase 4:**

Objective: Utilizing Linear Regression for house price prediction.

**Key Activities:**

* Loading the dataset and initial preprocessing steps.
* Feature selection and scaling.
* Selecting dependent and independent variables.
* Splitting data into training and testing sets.
* Training a Linear Regression model and evaluating its performance.

**Dataset Description:**

The dataset comprises essential information related to house features and prices. It contains the following features:

**Avg. Area Income:**

The average income of residents in the area where the house is located.

**Avg. Area House Age:**

The average age of houses in the area.

**Avg. Area Number of Rooms:**

The average number of rooms in houses in the locality.

**Avg. Area Number of Bedrooms**:

The average number of bedrooms in houses in the area.

**Area Population**:

The population of the area where the house is situated.

**Address:**

The address of the house.

**Price:**

The target variable representing the price of the house.

**Data Preprocessing Steps**

**1. Data Cleaning:**

**Handling Missing Values:**

Checked for missing data in the dataset and confirmed that there were no missing values present across features.

**Addressing Irrelevant Data**:

Removed the 'Address' column, which was considered non-impactful for predicting house prices.

**2. Feature Selection:**

**Correlation Analysis:**

Calculated the correlation matrix and visualized it using a heatmap. Identified correlations between features and the target variable (house price). Features with stronger correlations were given higher priority.

**Feature Importance:**

Utilized various techniques to understand the importance of features in predicting house prices.

**3. Feature Scaling:**

Utilized StandardScaler to scale the data, ensuring that all features were on a comparable scale. This step is particularly important for various machine learning algorithms to perform optimally.

**4. Data Transformation:**

If categorical variables were present in the dataset, conversion to numerical representation (label encoding or one-hot encoding) would be essential for compatibility with machine learning algorithms.

**Model Training**

**Data Splitting:**

Split the data into training and testing sets using a 70-30 split ratio.

**Method Used:**

The train\_test\_split function from the scikit-learn library was applied to split the dataset into two parts: the training set and the testing set.

**Linear Regression:**

Utilized the LinearRegression model from the scikit-learn library for training.

**Regression Algorithm Choice:**

The selected regression algorithm for this project is Linear Regression.

**Why Linear Regression ?**

**Simplicity and Interpretability:**

Linear Regression is a straightforward and easy-to-understand algorithm that assumes a linear relationship between the input features and the target variable.

**Efficiency:**

It serves as a good starting point for regression tasks, especially when exploring relationships between multiple variables.

**Model Interpretability:**

Linear Regression allows us to interpret the impact of each feature on the target variable through the coefficients.

**Evaluation Metrics:**

The project uses various evaluation metrics to assess the model's performance:

**Why these Metrics?**

Comprehensiveness:

Utilizing multiple evaluation metrics offers a more comprehensive assessment of the model's performance. Each metric addresses distinct aspects, providing a holistic understanding of how the model behaves in different scenarios or facets.

Different Aspects:

* R² Score : Indicates the model's capability to capture the variance in the data.
* Mean Absolute Error : Provides insight into the magnitude of errors.
* Mean Squared Error : Emphasizes larger errors more due to squaring the differences.

Balancing Metrics:

Evaluating multiple metrics helps in making a more balanced judgment about the model's performance, considering both the accuracy and variance explanations.

* R² measures the variance explanation capacity, indicating how well the model explains the variance in the target.
* MAE and MSE concentrate on the accuracy and types of errors made by the model, offering insights into how the model performs in predicting actual values.