**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 Process:**

**Data Source Selection:** Choose a reliable dataset with relevant house features and prices.

**Data Preprocessing:** Clean data, handle missing values, remove outliers, and convert categorical features to numerical representations.

Feature Selection: Analyze feature correlations and identify influential features.

**Model Selection**: Choose between Linear Regression and Random Forest Regressor based on dataset size and interpretability requirements.

**Model Training:** Divide the dataset into training and testing sets and train the selected model.

**Evaluation:** Employ metrics like MAE, RMSE, and R-squared to evaluate the model's effectiveness.

**Phases of Development:**

**Phase1**: Utilize machine learning to predict house prices.

Steps: Data Source Selection, Preprocessing, Feature Selection, Model Selection, Training, and Evaluation.

**Phase2**: Enhance the model using deep learning techniques.

Actions: Explore, preprocess, design a neural network, train, and evaluate.

**Phase3**: Load the dataset and preprocess for analysis.

Key Steps: Implement the Design by coding the predictive model, import data, libraries, handle missing values, feature selection, and scaling.

**Phase4**: Utilize Linear Regression for house price prediction.

Activities: Load data, preprocess, select features, split data, train a Linear Regression model, and evaluate.

**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.