Phase 1: Problem Definition and Design Thinking

Problem Definition :

The problem at hand is to predict house prices using machine learning techniques. The main objective is to develop a model that can accurately predict 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 several key steps, including data preprocessing, feature engineering, model selection, training, and evaluation.

Design Thinking :

1. Data Source :

   The first step in solving this problem is to choose an appropriate dataset that contains information about houses and their associated features. The dataset should include features like location, square footage, number of bedrooms, number of bathrooms, and the corresponding price of the houses. For this project, we will use the dataset available at the following link: [USA Housing Dataset](https://www.kaggle.com/datasets/vedavyasv/usa-housing).

2. Data Preprocessing :

   Data preprocessing is a crucial step to ensure that the dataset is suitable for training a machine learning model. This step involves the following tasks:

   1. Data Cleaning:
      Check for and handle any missing or erroneous data points in the dataset. This may include filling in missing values or removing outliers.
   2. Feature Scaling:
      Scale numerical features if they have different units or ranges to ensure that they contribute equally during model training.
   3. Feature Encoding:
      Convert categorical features, such as location, into numerical representations using techniques like one-hot encoding or label encoding.

3. Feature Selection :

   Not all features in the dataset may be equally important for predicting house prices. Feature selection is the process of choosing the most relevant features that have the most significant impact on the target variable (house prices). This can be done through techniques such as correlation analysis or feature importance scores from the selected machine learning algorithm.

4. Model Selection :
   Selecting the right machine learning algorithm for regression is critical. In this phase, we will explore various regression algorithms and choose the one that best fits the problem. Some of the potential algorithms to consider include:

   -Linear Regression: A simple and interpretable model.

   -Random Forest Regressor: A more complex model that can capture non-linear relationships.

   The choice of the algorithm will depend on the dataset and the performance metrics.

5. Model Training :

   Once we have chosen the regression algorithm, we will proceed with training the model on the preprocessed dataset. This involves splitting the data into training and testing sets, fitting the model to the training data, and then assessing its performance on the testing data.

6. Evaluation :

   To measure the performance of our model, we will use various regression metrics, including:

   - Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual prices.

   - Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between predicted and actual prices.

   - R-squared (R2): Indicates the proportion of the variance in the target variable that is predictable from the independent variables.

   We will aim to minimize MAE and RMSE while maximizing R-squared to build an accurate predictive model.