Problem Definition and Design Thinking:

Problem Definition:

The problem at hand is to predict house prices using machine learning techniques. The primary objective is to develop a model that accurately predicts house prices based on various 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, model training, and model evaluation.

Design Thinking Approach:

1. Data Source Selection:

- ❖ Begin by selecting a suitable dataset that contains comprehensive information about houses. The provided dataset at [Dataset Link](https://www.kaggle.com/datasets/vedavyasv/usa-housing) appears to be a relevant choice, as it includes features like location, square footage, bedrooms, bathrooms, and house prices.

2. Data Preprocessing:

- ❖ Cleanse and preprocess the dataset to ensure data quality and consistency.

- Handle missing values using appropriate techniques such as imputation (mean, median) or data removal, depending on the nature of the missing data.
- Convert categorical features, like location, into numerical representations through methods like one-hot encoding or label encoding.
- Normalize or standardize numerical features to bring them to a consistent scale, which can enhance model performance.

3. Feature Selection:

- Identify the most relevant features that have a significant impact on house prices.
- Employ techniques such as correlation analysis, feature importance from tree-based models, or domain expertise to select the most influential predictors.
- Eliminate redundant or irrelevant features to streamline model training.

4. Model Selection:

- Choose a suitable regression algorithm based on the problem's characteristics. Common options include:
- Linear Regression: Simple and interpretable, suitable as a baseline model.
- Random Forest Regressor: Robust and capable of capturing complex relationships.
- Gradient Boosting Regressors: Can provide high predictive accuracy.
- Experiment with multiple algorithms to identify the one that performs best for this specific problem.

5. Model Training:

- ❖ Split the preprocessed dataset into training and testing subsets to evaluate model performance.
- ❖ Train the selected regression model on the training data, adjusting hyperparameters as needed.
- ❖ Implement cross-validation techniques, such as k-fold cross-validation, to assess model stability and reliability.

6. Evaluation:

- ❖ Assess the model's performance using key regression metrics, including:
- ❖ Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual prices.
- ❖ Root Mean Squared Error (RMSE): Quantifies the standard deviation of prediction errors.
- ❖ R-squared (R2): Indicates the proportion of variance in the target variable explained by the model.
- ❖ Compare the model's results against baseline models and industry standards to gauge its effectiveness.

- ➢ Throughout this process, maintain a structured workflow, document all decisions and procedures, and iterate as necessary to fine-tune the model's performance. Regularly communicate with stakeholders to ensure alignment with project objectives and expectations.