Phase 1

Problem Definition and Design Thinking

Your problem definition and design thinking process for predicting house prices using machine learning techniques are well-structured. Here's a more detailed breakdown of the steps you've outlined:

1. Data Source:

- Identify and obtain a dataset that contains relevant information about houses. You may consider sources like real estate websites, government databases, or publicly available datasets.

2. Data Preprocessing:

- Data Cleaning: Clean the dataset by handling missing values, duplicates, and outliers. This ensures that your model is trained on high-quality data.
- Data Transformation: Convert categorical features into numerical representations using techniques like one-hot encoding or label encoding.
- Feature Scaling: Scale numerical features if needed, to bring them to a similar range. Common techniques include Min-Max scaling or standardization.

3. Feature Selection:

- Use techniques like correlation analysis, feature importance from tree-based models, or domain knowledge to select the most relevant features for predicting house prices. Eliminate irrelevant or redundant features to simplify the model.

4. Model Selection:

- Choose an appropriate regression algorithm. Some popular options for house price prediction include:
  - Linear Regression
  - Decision Trees
  - Random Forest Regressor
  - Gradient Boosting Regressor
  - Support Vector Regression (SVR)
- Experiment with different algorithms to find the one that works best for your dataset.

5. Model Training:

- Split the dataset into training and testing sets to evaluate your model's performance.

- Train your selected regression model on the training data using appropriate hyperparameters and cross-validation techniques to avoid overfitting.

6. Evaluation:

- Evaluate your model's performance using various regression metrics such as:
- **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 difference between predicted and actual prices, giving more weight to larger errors.
- **R-squared (R2):** Measures the proportion of the variance in the target variable explained by the model. Higher values indicate better fit.
- Visualize the results through scatter plots, residual plots, and other visualization techniques to gain insights into model performance.

7. Iterate and Improve:

- If the initial model performance is not satisfactory, iterate through the process by experimenting with different features, models, and hyperparameters.
- Fine-tune your model based on the insights gained during evaluation.

8. Deployment:

Once you have a well-performing model, you can deploy it to make predictions on new data, such as predicting house prices for real-time or future listings.

Throughout the process, document your decisions, experiments, and results to ensure transparency and repeatability. Continuous monitoring and updates to the model may also be necessary to account for changing market conditions or data shifts.