PHASE 1: PREDICTION OF HOUSE PRICES

1. Data Collection:

- Gather a comprehensive dataset that includes relevant information about houses. This could encompass data sources such as real estate listings, property tax records, or publicly available datasets.
- Ensure that your dataset has a variety of features, including both numerical (e.g., square footage, number of bedrooms) and categorical (e.g., location, type of house) variables.
- Pay attention to data quality, as missing or inaccurate data can significantly impact the performance of your model.

2. Data Pre-processing:

➢ Data Cleaning:

- Identify and handle missing data: You can choose to impute missing values with averages or medians for numerical features and use mode for categorical features, or you may decide to remove rows or columns with excessive missing data.

- Outlier detection and treatment: Identify outliers in the data and decide whether to remove them, transform them, or leave them as-is based on domain knowledge.
- Error correction: Check for data entry errors and correct them if necessary.

➢ Feature Scaling and Transformation:

- Standardization: Standardize numerical features to have a mean of 0 and a standard deviation of 1. This is important for algorithms sensitive to feature scales, like gradient descent-based methods.
- Normalization: Normalize features to a specific range, like [0, 1], if needed.
- Log transformation: Apply log transformations to features that exhibit skewed distributions, which can help improve model performance.

➢ Categorical Encoding:

- One-Hot Encoding: Convert categorical variables into binary vectors, where each category becomes a binary feature.
- Label Encoding: Assign a unique integer to each category. This is suitable for ordinal categorical data.

3. Feature Engineering:

- ➢ Creating New Features:
    - Generate new features that may be more informative, such as the age of the house (current year minus year built).
    - Calculate ratios or proportions between features, like the price per square foot.
- ➢ Feature Selection:
    - Utilize techniques like correlation analysis to identify relationships between features and the target variable.
    - Employ feature importance scores from tree-based models like Random Forests or Gradient Boosting to select the most relevant features.
- ➢ Model Selection:

- Choose appropriate algorithms for regression tasks:
- Linear Regression: Simple and interpretable but assumes a linear relationship between features and target.
- Decision Trees and Random Forests: Non-linear models that can capture complex relationships in the data.
- Gradient Boosting: Ensemble method that combines multiple weak learners to create a strong predictive model.
- Support Vector Machines (SVM): Effective for high-dimensional data.

- Neural Networks: Deep learning models that can capture intricate patterns but may require more data and computational resources.

5. Model Training:

- Split your dataset into training and testing sets to assess model performance.
- Train the selected model on the training data using the chosen algorithm. The model will adjust its parameters to learn patterns in the data.

6. Model Evaluation:

- Use regression evaluation metrics:
- Mean Absolute Error (MAE): Measures the average absolute difference between predicted and actual values.
- Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.
- Root Mean Squared Error (RMSE): RMSE is the square root of MSE and provides a more interpretable measure.
- Perform cross-validation to assess how well the model generalizes to new data and to detect overfitting.

7. Hyper parameter Tuning:

- Experiment with different hyper parameter settings to optimize model performance. You can use techniques like grid search or random search to find the best combination of hyperparameters.

8. Deployment:

- .Once you have a well-performing model, deploy it as a service or integrate it into a web application where users can input house features and get price predictions

9. Monitoring and Maintenance:

- Continuously monitor the model's performance in a real-world setting and retrain it periodically with new data to keep it up to date.

10. Interpretability:

- Use techniques like feature importance analysis to understand which features are most influential in making predictions. This can provide valuable insights for stakeholders.
- Remember that the success of the house price prediction model depends on the quality of the data, the appropriateness of chosen techniques, and careful evaluation and refinement. It's often an iterative process

where you may need to revisit and improve various steps to achieve the best results.