**PHASE I**

**House Price Prediction Project Proposal**

**SREE VARSHINI S**

**2021503563**

**ARTIFICIAL INTELLIGENCE**

**NAAN MUDHALVAN**

**Problem Statement**

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

**Design Thinking**

1. **Data Source**

For this project, we will use the "USA Housing" dataset, which is available on Kaggle. This dataset contains information about houses in the USA, including features like location, square footage, bedrooms, bathrooms, and price. The dataset can be accessed here: USA Housing Dataset.

1. **Data Preprocessing**

Data preprocessing is a critical step in any machine learning project. We will perform the following steps:

1. Data Cleaning: Check for and handle any missing or duplicate data points.
2. Feature Engineering: Create new features if necessary. For example, we can calculate the price per square foot, which might be a useful feature.
3. Handling Categorical Data: Convert categorical features such as location into numerical representations using techniques like one-hot encoding or label encoding.
4. Scaling Features: Scale numerical features to ensure they have the same range. This can help improve the performance of certain machine learning algorithms.
5. **Feature Selection**

Not all features may be equally important for predicting house prices. We will use techniques such as feature importance scores and correlation analysis to select the most relevant features. This will help in improving model performance and reducing dimensionality if necessary.

1. **Model Selection**

The choice of the regression algorithm is crucial for this problem. We will consider several regression algorithms, including but not limited to:

Linear Regression: A simple and interpretable model that can serve as a baseline.

Random Forest Regressor: An ensemble method that can capture complex relationships in the data.

Gradient Boosting Regressor: Another ensemble method that can handle non-linear relationships.

We will evaluate the performance of these models and choose the one that provides the best results.

1. **Model Training**

Once the model is selected, we will train it using the preprocessed data. We will split the dataset into training and testing sets to assess the model's performance on unseen data.

1. **Evaluation**

We will evaluate the model's performance using the following regression metrics:

Mean Absolute Error (MAE): This measures the average absolute difference between predicted and actual house prices.

Root Mean Squared Error (RMSE): This measures the square root of the average squared difference between predicted and actual prices. It penalizes larger errors more heavily than MAE.

R-squared (R²): This provides an indication of how well the model explains the variance in the data. A higher R² value indicates a better fit.

**Conclusion**

Summarize the results, discuss the model's performance, and suggest potential future improvements or extensions to the project.