**PREDICTING HOUSE PRICES USING MACHINE LEARNING**

**OBJECTIVE:**

The objective of this project is to predict house prices accurately based on location, number of rooms and bathrooms, square footage, and other relevant factors.

**PHASE 1: PROBLEM DEFINITION AND DESIGN THINKING**

**PROBLEM DEFINITION:**

Accurately estimating the value of a real estate is an important problem for many Stakeholders including house owners, house buyers, Agents, creditors, and investors. It is also a difficult one. Though it is common knowledge that factors Such as size, number of Rooms, and location affect the price, there are many other things at Play. Additionally, prices are sensitive to changes in market demand And the Peculiarities of each situation, such as when a property needs to be urgently sold. The Sales price of a property can be predicted in various ways but is often based on Regression Techniques. All regression techniques essentially involve one or more Predictor variables as input and a single target variable as the output. In In this paper, we compare different machine-learning method's Performance in predicting the selling price of houses based on Several features such as the area, the number of beds- and Bathrooms, and the geographical position.

**DESIGN THINKING:**

1. **DATA SOURCE:**

A good data source for house price prediction using machine Learning should be Accurate and complete, covering the geographic Area of interest, and Accessible.

**Dataset Link:**

(<https://www.kaggle.com/datasets/vedavyasv/usa-housing>)

**2.** **DATA PREPROCESSING:**

**Handle Missing Values:**

Identify and handle missing data in the dataset. You can fill in missing Values with the mean, median, or mode of the respective feature, or Consider more advanced imputation methods.

**Remove Duplicates:**

Check for and remove duplicate entries if they exist.

**Handling Categorical Variables:**

One-Hot Encoding: Convert categorical variables (e.g., neighborhood Names) into binary (0/1) columns for each category using one-hot Encoding.

Label Encoding: For ordinal categorical variables (e.g., “low,” “Medium,” “high”), you can use label encoding to convert them into Numerical values.

**Normalization/Scaling:**

Scale numerical features: Use techniques like Min-Max scaling or Z-Score.

Normalization to ensure that all numerical features have similar Scales.

**3. FEATURE SELECTION:**

Correlation-based feature selection: Identify features that are

Highly correlated with the target variable (house prices)

Using correlation coefficients. Feature features with strong

Correlations.

Recursive feature elimination (RFE): This technique starts with all Features and then Recursively removes the least important feature until a desired number of features remains. The importance of a Feature is measured using a variety of methods, such as Cross-Validation or the coefficient of determination (R-squared).

**4. MODEL SELECTION:**

Model selection is the process of choosing a suitable machine-learning algorithm for a given machine-learning task. The goal of model selection is to find an algorithm that is both accurate and efficient. There are a variety of machine learning algorithms that can be used for house price Prediction.  Some of the most common algorithms include:

 Linear regression: Linear regression is a simple but effective algorithm for house price Prediction. Linear regression models the relationship between the house price and the Features using a linear function.

Random forest regression: Random forest regression is a more complex algorithm that builds a multitude of decision trees to predict the house price. Random forest regresses are typically more accurate than linear regression models, but they can be more computationally expensive to train.

Gradient boosting regression: Gradient boosting regression is another complex algorithm. That builds a sequence of decision trees to predict the house price. Gradient boosting regressions are typically more accurate than random forest regresses, but they can be even more computationally expensive to train.

Choose the Linear Regression algorithm as your predictive model, given that you want to predict house prices, a continuous target variable.

**5. MODEL TRAINING:**

Use the training data to fit the Linear Regression model. This involves finding the best-fitting line (or hyperplane in the case of multiple features) that minimizes the sum of Squared differences between predicted and actual house prices.

**6. Evaluation Matrices:**

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted values And the actual target values. It provides insight into the average magnitude of errors made by The model.

Root Mean Squared Error (RMSE): RMSE is another common metric that calculates the square root of the average squared differences between predicted and actual values. It provides information about the  Typical magnitude of errors and gives higher penalties for larger errors.

S-squared (R2): T-R-squared quantifies the proportion of the variance in the target  Variable that is explained by the model. It ranges from 0 to 1, where a higher value  Indicates a better fit. It is often used to assess how well the model captures the  Variation in the data.

 These metrics quantify the error between predicted and actual prices.