PHASE 5

Our project was dedicated to "Predicting House Prices using Machine Learning." The primary goal was to develop a machine learning model capable of accurately forecasting house prices based on various attributes and features.

Design Thinking Process:

Understanding the Problem: Our journey commenced with a profound comprehension of the real-world significance of accurate house price predictions.

Data Collection: The first step involved using the "usa-housing" dataset from Kaggle. This dataset encompassed vital attributes essential for house price prediction, including factors such as square footage, bedroom count, and geographical location.

Data Preprocessing: This phase encompassed the preparation of the dataset to build machine learning models. Tasks included addressing missing data, encoding categorical variables, and segregating the dataset into distinct training and testing subsets.We also found correlations between each attribute with that of the target variable "Price".

Model Selection: We decided on using two models and chose the one which proved to perform better. In our case, this was the linear regression model (accuracy: 91%) which performed better than the Random Forest model (accuracy: 88%).

Model Training: The model training phase involved acquainting the Linear Regression model with the dataset. This allowed it to understand the relationships between the various features and the target variable, in this case, house prices.

Evaluation: The project's success rested on the performance of the model. To gauge this, we employed specific evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), to quantify accuracy and reliability.

Improvement: The iterative nature of machine learning projects was embraced as we continually refined the model. This involved adjusting algorithm parameters, considering alternative models, and engaging in feature engineering to enhance predictive performance.

Phases of Development:

Data Collection: The first step involved using the "usa-housing" dataset. This dataset housed a comprehensive set of features and corresponding house prices.

Data Preprocessing:

- Missing data was methodically addressed.
- Categorical variables were encoded, rendering them compatible with the chosen algorithms.
- The dataset was divided into separate training and testing segments for model training and evaluation..

Model Selection: The selection of the Linear Regression algorithm was made purposefully, given its inherent simplicity and interpretability. A Random forest model was also

Model Training: The model was trained using the training dataset, enabling it to absorb the intricate relationships between features and the target variable, namely house prices.

Model Evaluation: The effectiveness of the model was appraised using established regression evaluation metrics. Common metrics, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE) were deployed to assess accuracy and reliability.

Choice of Regression Algorithm and Evaluation Metrics:

Regression Algorithm (Linear Regression): The choice of Linear Regression was a deliberate one, stemming from its capacity to predict numerical values accurately, making it an apt selection for predicting house prices.

Evaluation Metrics: The metrics selected for evaluation encompassed:

Mean Absolute Error (MAE): MAE quantified the average absolute disparity between predicted and actual house prices, offering a clear measure of predictive accuracy.
Mean Squared Error (MSE): MSE computed the mean of squared discrepancies between predicted and actual values, with a focus on larger errors and their implications.

During the course of our project, we ventured into the realm of Random Forest, an alternative model. However, we observed that while it possessed power, it yielded an accuracy of 88%, falling slightly behind our Linear Regression model, which yielded an accuracy rate of 91%. This comparison provided the necessary information to choose our final model.

Our project adhered to a systematic approach, progressing from problem comprehension to model selection and rigorous evaluation. The usage of Linear Regression and Random Forest models provided valuable insights for model selection.