**PREDICTING HOUSE PRICE USING MACHINE LEARNING**

**Phase-2 Document Submission**

**1.Data Source**

A good data source for house price prediction using machine learning should be Accurate, Complete, Covering the geographic area of interest, Accessible.

**Dataset Link: https://www.kaggle.com/datasets/vedavyasv/usa-housing**

**2.Data Preprocessing**

Data preprocessing is the critical first step in any machine learning project. It involves cleaning the data, removing outliers, and handling missing values to prepare the dataset for model training. In the context of the house price prediction project, let's elaborate on the specific steps:

**a) Duplicate Removal:**

Duplicate rows can introduce bias into the model. We will identify and remove duplicates, typically by sorting the dataset based on a unique identifier (e.g., property ID) and then eliminating consecutive rows with the same identifier.

**b) Handling Missing Values:**

Missing data is common and needs to be addressed. We will utilize suitable methods such as:

➢ Mean Imputation: Replace missing values with the mean of the feature for the remaining rows. This is appropriate for numerical features.

➢ Median Imputation: If data contains outliers, median imputation can be more robusas it is less sensitive to extreme values.

**c) Categorical Variable Encoding:**

Categorical variables, such as property type or location, need to be converted into numerical form so that machine learning models can process them. Two common approaches include:

➢ One-Hot Encoding: Create binary columns for each category, representing the presence or absence of that category.

➢ Label Encoding: Assign a unique integer to each category, preserving the ordinal relationship if applicable.

**d) Data Normalization:**

To ensure that all features are on a consistent scale, normalization techniques can be applied. This includes:

➢ Standardization: Scaling features to have a mean of 0 and a standard deviation of 1. ➢ Min-Max Scaling: Scaling features to a specified range (e.g., 0 to 1)

**CONCLUSION:**

In Phase 2, we have established a clear understanding of our goal: to predict house prices using machine learning. We outlined a structured approach that includes data source selection, data preprocessing, feature selection, model selection, model training, and evaluation. This sets the stage for our project's successful execution in subsequent phases.