PREDICTING HOUSE PRICES USING MACHINE LEARNING

|  |
| --- |
| **AI\_PHASE\_3** |
| NM:PROJECT |
|  |
| |  |  | | --- | --- | | **NAME** | **VISHVAJITH V** | | **REGESTER NUMBER** | **61772221T310** | | **DATE** | **26 - 10 - 2023** | |
|  |
| **COLLEGE : GCE-SALEM** |
| **26-Oct-23** |
|  |

## PHASE 3 : DEVELOPMENT PART 1

# INTRODUCTION :

* Machine learning involves training a computer to recognize patterns and make predictions based on data.
* In the case of house price prediction, we can use historical data on various features of a house, such as its location, size, and amenities, to train a machine learning model.
* Once the model is trained, it can analyze new data on a given house and make a prediction of its market value.

# DATASET LINK :

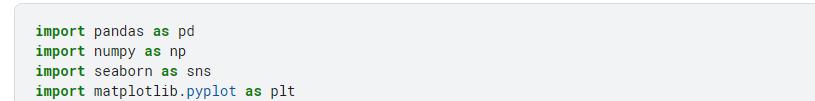
<https://www.kaggle.com/datasets/vedavyasv/usa-housing>

# PROBLEM STATEMENT :

* The problem 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.

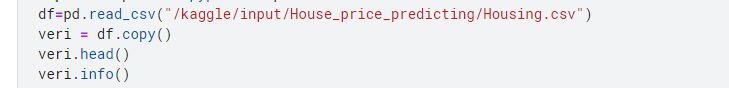
# IMPORTING NECESSARY LIBRARIES :

* pandas is used for data manipulation and loading.
* numpy is used for numerical operations.
* matplotlib and seaborn are used for data visualization.



# LOAD THE DATASET:

* We use the pandas library to import a dataset named 'Housing.csv' by employing the pd.read\_csv() function.
* The loaded data is stored in a data frame called df, which is a two dimensional table for data storage and manipulation.



# ALGORITHM :

* STEP 1 : Import the required libraries and modules, including pandas for data manipulation, scikit-learn for machine learning algorithms, and Linear Regression for the linear regression model.
  + STEP 2 : Loading the required dataset with pd.read\_csv and select the features we want to use for prediction (e.g., bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, and zip code), as well as the target variable (price).
  + STEP 3 : Split the data into a training set and a test set using the train\_test\_split function, with 80% of the data used for training and 20% for testing.
  + STEP 4: Create an instance of the linear regression model using Linear Regression(). We then perform the model training by calling the function fit() with the training data.
  + STEP 5: Demonstrate how to predict the price of a new house by creating a new data frame new house with the features of the house. We pass this data frame to the model's prediction function to obtain the predicted price.

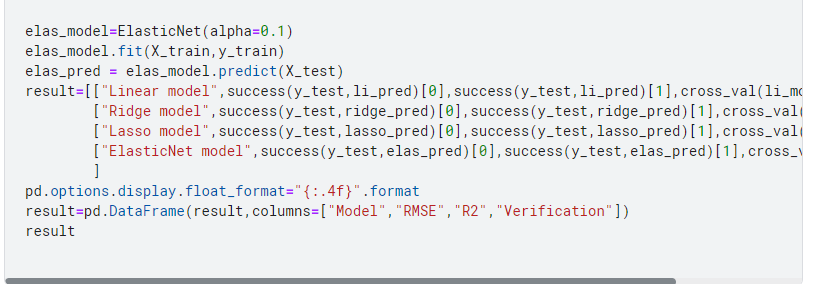
# Data Analysis :

* + EDA refers to the deep analysis of data so as to discover different patterns and spot anomalies.
  + Before making inferences from data it is essential to examine all your variables.
  + So here let’s make a heatmap using seaborn library.

# CODE :







#End of Program

# Data Cleaning :

* + Data Cleaning is the way to improvise the data or remove incorrect, corrupted or irrelevant data.
  + As in our dataset, there are some columns that are not important and irrelevant for the model training. So, we can drop that column before training.
  + There are 2 approaches to dealing with empty/null values
* We can easily delete the column/row (if the feature or record is not much important).
* Filling the empty slots with mean/mode/0/NA/etc. (depending on the dataset requirement). As Id Column will not be participating in any prediction. So we can Drop it.

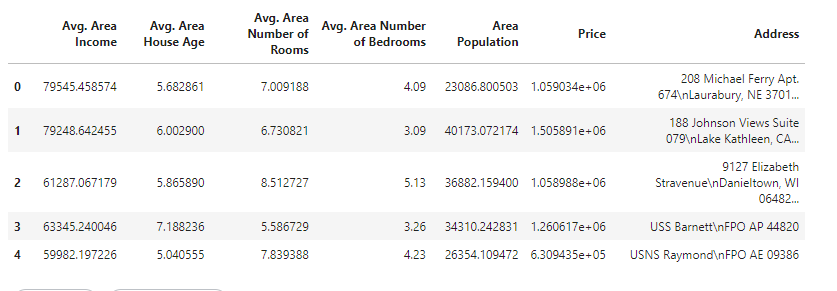
# Model and Accuracy :

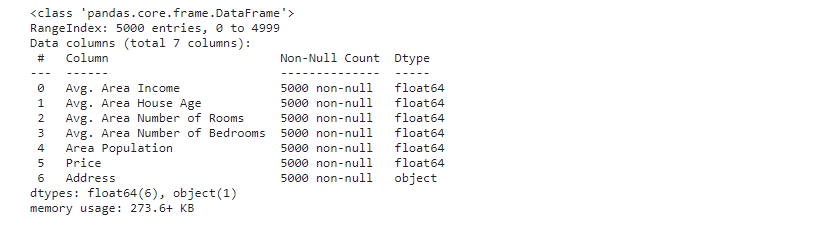
* As we have to train the model to determine the continuous values, so we will be using these regression models.
* 1. SVM-Support Vector Machine
* 2. Random Forest Regressor
* 3. Linear Regressor
* And To calculate loss we will be using the mean\_absolute\_percentage\_error module. • It can easily be imported by using sklearn library.
* SVM – Support vector Machine -> SVM can be used for both regression and classification model. It finds the hyperplane in the n-dimensional plane. To read more about svm refer this.

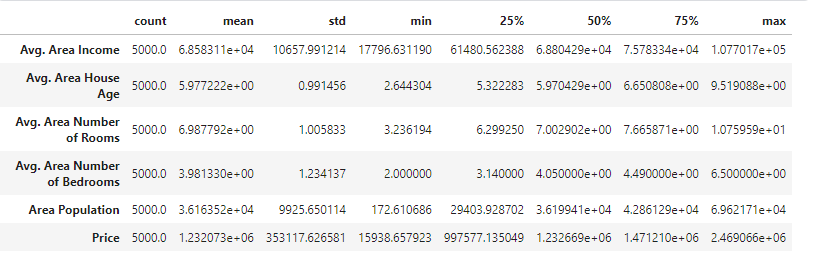
# Label categorical features :

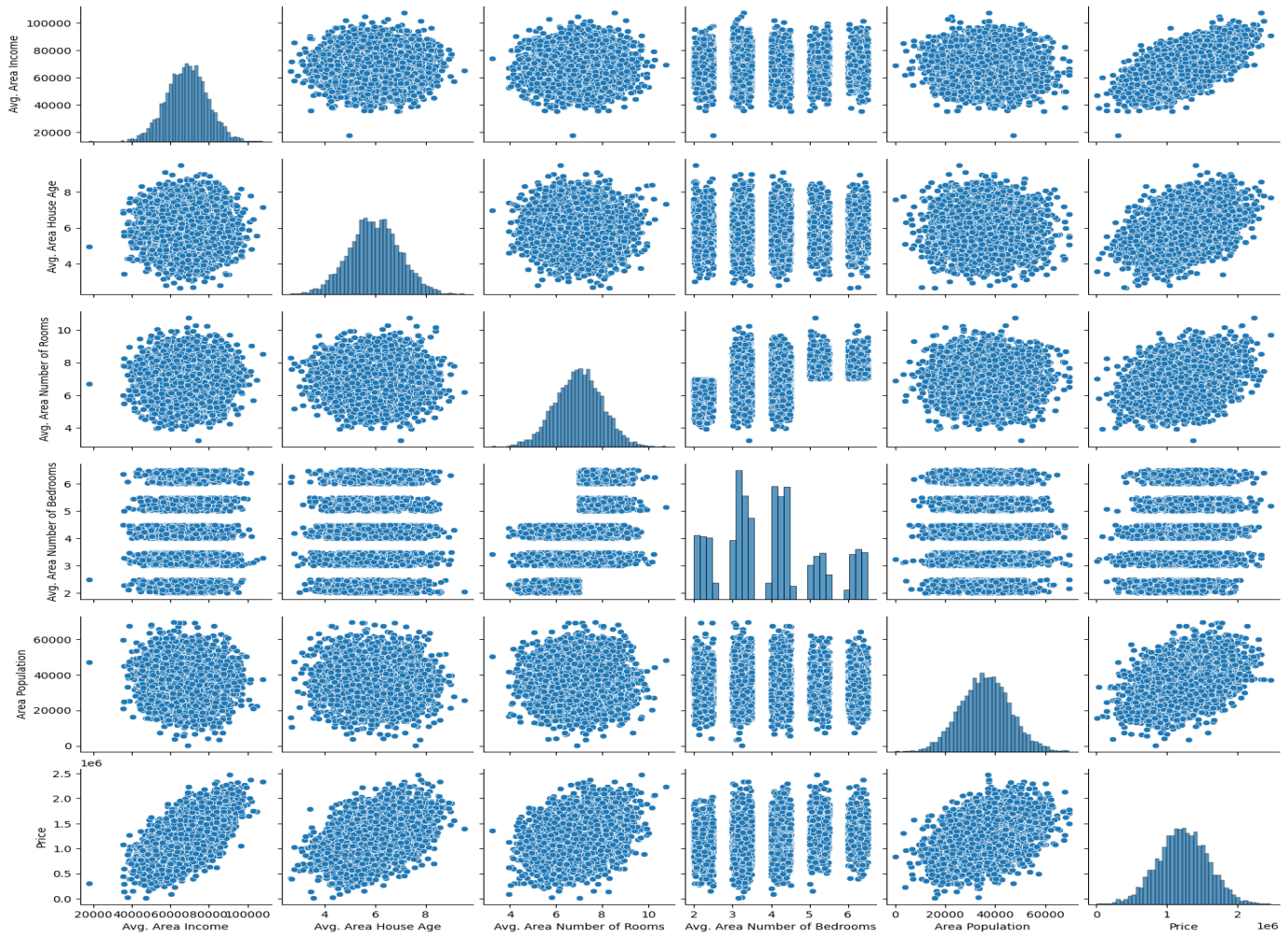
* + One hot Encoding is the best way to convert categorical data into binary vectors.
  + This maps the values to integer values. By using One Hot Encoder, we can easily convert object data into int.
  + So for that, firstly we have to collect all the features which have the object datatype. To do so, we will make a loop.

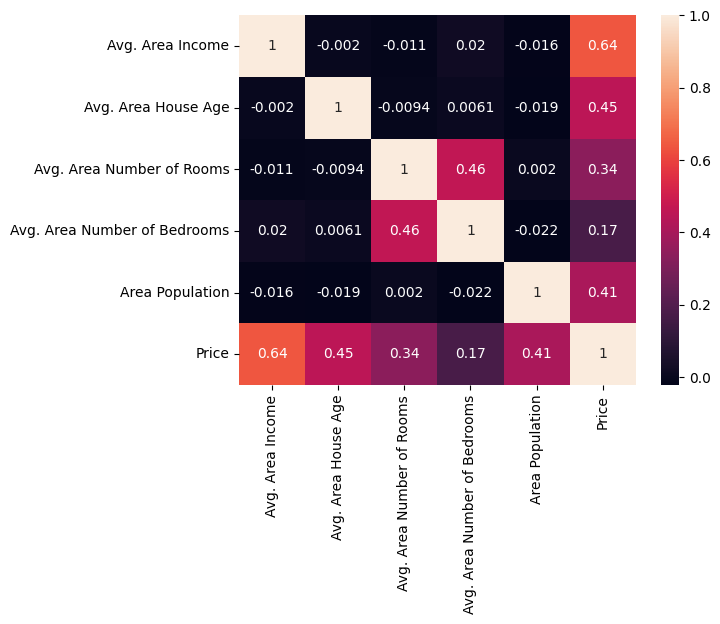
# OUTPUT :

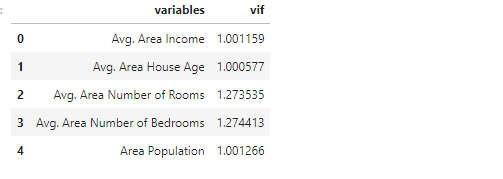


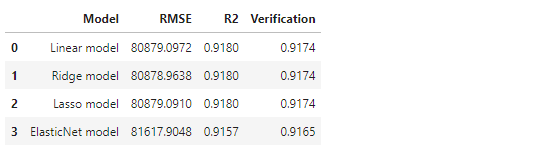












# CONCLUSION :

* In conclusion, using machine learning in Python is a powerful tool for predicting house prices.
* By gathering and cleaning data, visualizing patterns, and training and evaluating our models, we can make informed decisions in the dynamic world of real estate.
* By leveraging advanced algorithms and data analysis, we can make accurate predictions and inform decision-making processes.
* This approach empowers buyers, sellers, and investors to make informed choices in a dynamic and competitive market, ultimately maximizing their opportunities and outcomes.