PREDICTING HOUSE PRICE USING MACHINE LEARNING

Phase – 3

We started building our project by loading the dataset, Data processing and exploratory data analysis in Google colab Notebook.

Loading the Dataset:

Downloaded the **USA_Housing.csv** dataset from the Kaggle

We loaded a train.csv dataset using the **pandas** library.

```
import pandas as pd
dataset = pd.read_csv("USA_Housing.csv")
print(dataset.head())
```

```
   Avg. Area Income  Avg. Area House Age  Avg. Area Number of Rooms  \
0      79545.458574             5.682861                   7.009188
1      79248.642455             6.002900                   6.730821
2      61287.067179             5.865890                   8.512727
3      63345.240046             7.188236                   5.586729
4      59982.197226             5.040555                   7.839388

   Avg. Area Number of Bedrooms  Area Population         Price  \
0                          4.09     23086.800503  1.059034e+06
1                          3.09     40173.072174  1.505891e+06
2                          5.13     36882.159400  1.058988e+06
3                          3.26     34310.242831  1.260617e+06
4                          4.23     26354.109472  6.309435e+05

                                             Address
0  208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1  188 Johnson Views Suite 079\nLake Kathleen, CA...
2  9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3                       USS Barnett\nFPO AP 44820
4                      USNS Raymond\nFPO AE 09386
```

+ Code + Text

Data Processing:

Now, we categorize the features depending on their datatype (int, float, object) and then calculate the number of them.

```
dataset.shape
```

```
(5000, 7)
```

```
obj = (dataset.dtypes == 'object')
object_cols = list(obj[obj].index)
print("Categorical variables:",len(object_cols))

int_ = (dataset.dtypes == 'int')
num_cols = list(int_[int_].index)
print("Integer variables:",len(num_cols))

fl = (dataset.dtypes == 'float')
fl_cols = list(fl[fl].index)
print("Float variables:",len(fl_cols))
```

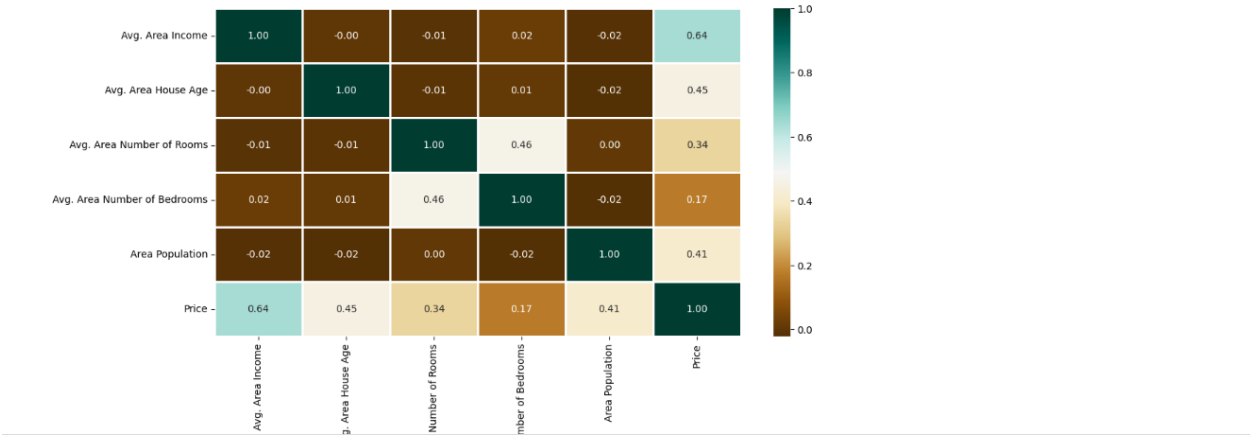
```
Categorical variables: 1
Integer variables: 0
Float variables: 6
```

Exploratory 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.Using Seaborn libraries.

```
plt.figure(figsize=(12, 6))
sns.heatmap(dataset.corr(),
        cmap = 'BrBG',
        fmt = '.2f',
        linewidths = 2,
        annot = True)
```

```
<ipython-input-26-8feaf9d49085>:2: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric
  sns.heatmap(dataset.corr(),
<Axes: >
```



Team Members:

K.Santhosh

J.Saravanakumar

M.Sangara Sequvar

C.Mohan

JP COLLEGE OF ENGINEERING