COLLEGE CODE: 9133

COURSE:Artificial intelligence

PHASE 5:

PROJECT TITLE: House price predictor

Team Members:

maalinivcse2021@gmail.com- Maalini.V

pavithrancse2021@gmail.com- Pavithra.N

divyadharshiniscse2021@gmail.com- Divyadharshini.S

manoranjithampcse2021@gmail.com- Manoranjitham.P

priyadharshiniscse2021@gmail.com- Priyadharshini.S

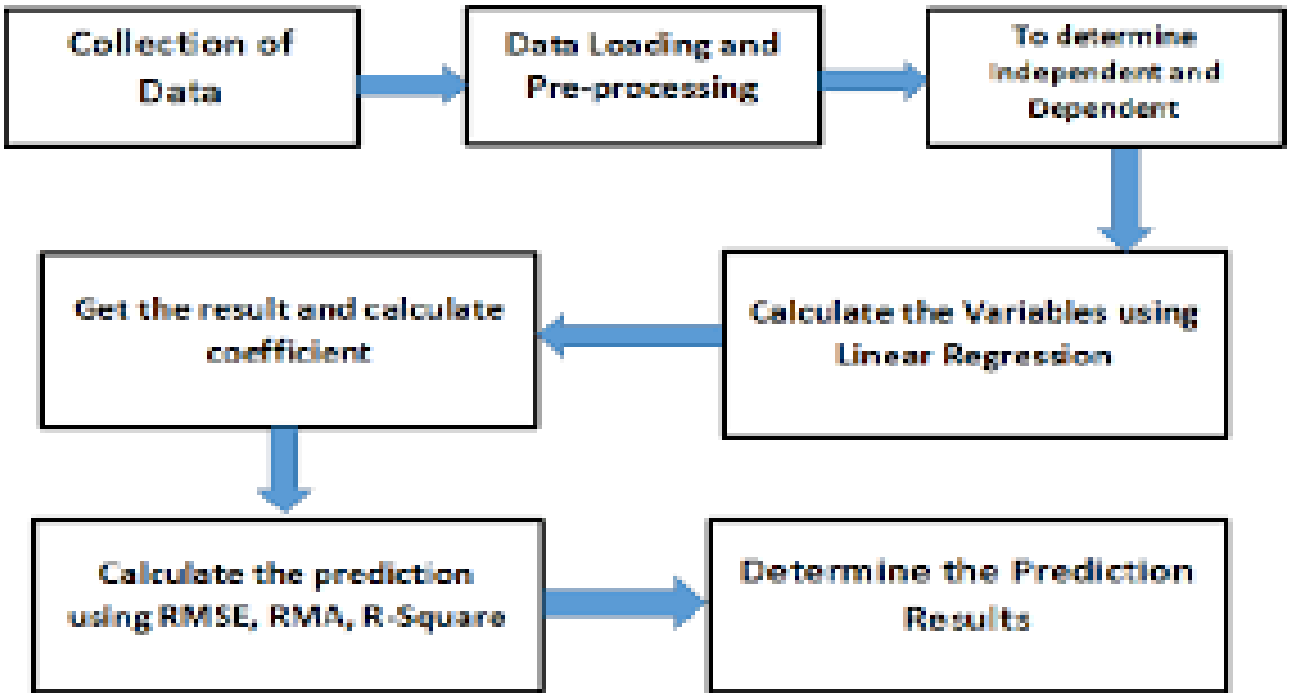
Problem statement:

To predict the house price using machine learning algorithm.

Design thinking process:

Gather a dataset containing historical information on houses. This dataset should include features like square footage, number of bedrooms and bathrooms, location, and any other relevant information. It should also include the actual selling prices of these houses.

Data preprocessing steps:



Dataset used: https://www.kaggle.com/datasets/vedavyasv/usa-housincontaining information about houses, including features like location, square footage, bedrooms, bathrooms, and price.

Data preprocessing steps:

Clean the data by handling missing values and outliers. Encode categorical features if necessary (e.g., converting locations into numerical values using one-hot encoding). Split the dataset into training and testing sets to evaluate the model later.

Feature Extraction:

Analyze feature importance and select the most relevant features. Create new features if they can provide valuable information (e.g., a "price per square foot" feature).

Machine learning algorithm:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Model training:

Train the selected model on the training data. The model learns patterns and relationships within the data during this process. The training process typically involves adjusting model parameters to minimize a loss function.

Model evaluation:

After training, you need to evaluate the model's performance on the testing data. Common evaluation metrics include accuracy, precision, recall, F1-score for classification, and mean squared error, R-squared for regression.

CODE:

[1]:
```
import pandas as pd
import numpy as np
```

IMPORTING DATA

[2]:
```
houses=pd.read_csv('../input/usa-housing/USA_Housing.csv')
```

[3]:
```
houses
```

[3]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanielltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820

```
houses.columns
```

```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Address'],
      dtype='object')
```

```
houses.head(7)
```

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanielltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386
5	80175.754159	4.988408	6.104512	4.04	26748.428425	1.068138e+06	06039 Jennifer Islands Apt. 443\nTracyport, KS...
6	64698.463428	6.025336	8.147760	3.41	60828.249085	1.502056e+06	4759 Daniel Shoals Suite 442\nNguyenburgh, CO ...

[4]:
```
houses.columns
```

[4]:
```
Index(['Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms',
       'Avg. Area Number of Bedrooms', 'Area Population', 'Price', 'Address'],
      dtype='object')
```

[5]:
```
houses.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Avg. Area Income              5000 non-null   float64
 1   Avg. Area House Age           5000 non-null   float64
 2   Avg. Area Number of Rooms     5000 non-null   float64
 3   Avg. Area Number of Bedrooms  5000 non-null   float64
 4   Area Population               5000 non-null   float64
 5   Price                         5000 non-null   float64
 6   Address                       5000 non-null   object
dtypes: float64(6), object(1)
memory usage: 273.6+ KB
```

+ Code + Markdown

[6]:
```
houses.isnull().sum()
```

[6]:
```
Avg. Area Income              0
Avg. Area House Age           0
Avg. Area Number of Rooms     0
Avg. Area Number of Bedrooms  0
```

[7]:
```
houses.describe(include='all')
```

[7]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000	5.000000e+03	5000
unique	NaN	NaN	NaN	NaN	NaN	NaN	5000
top	NaN	NaN	NaN	NaN	NaN	NaN	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
freq	NaN	NaN	NaN	NaN	NaN	NaN	1
mean	68583.108984	5.977222	6.987792	3.981330	36163.516039	1.232073e+06	NaN
std	10657.991214	0.991456	1.005833	1.234137	9925.650114	3.531176e+05	NaN
min	17796.631190	2.644304	3.236194	2.000000	172.610686	1.593866e+04	NaN
25%	61480.562388	5.322283	6.299250	3.140000	29403.928702	9.975771e+05	NaN
50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06	NaN
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06	NaN
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06	NaN

[8]:
```
houses["Address"].value_counts()
```

[8]:
```
208 Michael Ferry Apt. 674\nLaurabury, NE 37010-5101          1
314 Christopher Square Apt. 404\nLake Ronaldville, SD 42025   1
21042 Wilson Islands Suite 238\nFischerchester, MP 42425-4129 1
```


50%	68804.286404	5.970429	7.002902	4.050000	36199.406689	1.232669e+06	NaN
75%	75783.338666	6.650808	7.665871	4.490000	42861.290769	1.471210e+06	NaN
max	107701.748378	9.519088	10.759588	6.500000	69621.713378	2.469066e+06	NaN

[8]:
```
houses["Address"].value_counts()
```

[8]:
```
208 Michael Ferry Apt. 674\nLaurabury, NE 37010-5101          1
314 Christopher Square Apt. 404\nLake Ronaldville, SD 42025   1
21042 Wilson Islands Suite 238\nFischerchester, MP 42425-4129 1
Unit 8831 Box 5748\nDPO AE 73012-7314                         1
481 Kaitlin Mission Apt. 309\nJodystad, IA 16947               1
                                                              ..
054 Carter Crescent Suite 674\nGlennport, WA 11140             1
8460 Kathleen Mission Apt. 482\nPort Amytown, KY 72816         1
3737 Hartman Rue\nReneestad, ID 69250-7718                     1
3465 Latoya Well\nNelsonmouth, MI 55741-4287                   1
37778 George Ridges Apt. 509\nEast Holly, NV 29290-3595        1
Name: Address, Length: 5000, dtype: int64
```

[9]:
```
houses.Address.unique()
```

[9]:
```
array(['208 Michael Ferry Apt. 674\nLaurabury, NE 37010-5101',
       '188 Johnson Views Suite 079\nLake Kathleen, CA 48958',
       '9127 Elizabeth Stravenue\nDanieltown, WI 06482-3489', ...,
       '4215 Tracy Garden Suite 076\nJoshualand, VA 01707-9165',
       'USS Wallace\nFPO AE 73316',
       '37778 George Ridges Apt. 509\nEast Holly, NV 29290-3595'],
      dtype=object)
```

[10]:
```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

[11]:
```
houses.head()
```

[11]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

[10]:
```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

[11]:
```
houses.head()
```

[11]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price	Address
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06	208 Michael Ferry Apt. 674\nLaurabury, NE 3701...
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06	188 Johnson Views Suite 079\nLake Kathleen, CA...
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06	9127 Elizabeth Stravenue\nDanieltown, WI 06482...
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06	USS Barnett\nFPO AP 44820
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05	USNS Raymond\nFPO AE 09386

[28]:
```
from sklearn.metrics import r2_score,mean_absolute_error,mean_squared_error
```

" Test prediction evaluation"

[29]:
```
r_squared=r2_score(test_pred,y_test)
print("R2 Score:", r_squared)
```

```
R2 Score: 0.9046796597914799
```

[30]:
```
linearmodel.score(X_test,y_test)
```

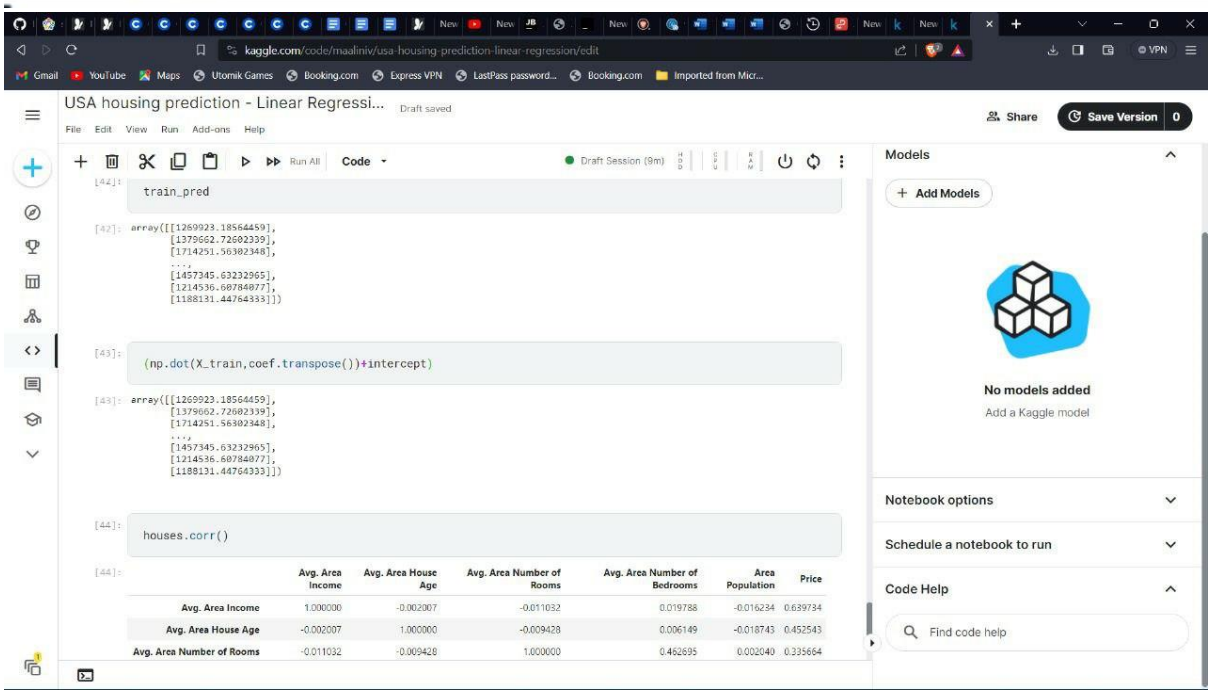
[30]: 0.9140423945227004

[31]:
```
mae=mean_absolute_error(y_test,test_pred)
print("Mean Absolute Error (MAE):", mae)
```

```
Mean Absolute Error (MAE): 81023.44047681554
```

```
(np.dot(X_test,coef.transpose())+intercept)
```

```
[40]: array([[1438564.40798403],
             [1161466.12056832],
             [1460026.18587627],
             [ 611038.86113306],
             [1387686.04009869],
             [ 513080.36497973],
             [1224117.61290795],
             [ 779130.42049552],
             [1212908.86013773],
             [1257728.14421955],
             [1107375.33113647],
             [1424933.36272296],
             [1880330.09111944],
             [1293467.06572774],
             [1227695.07801086],
             [1028919.62309891],
             [1324845.72239494],
             [1548918.57610611],
             [ 883897.17590746],
             [1096242.94629345],
             [1331480.18403404],
             [1584136.21941831],
             [1067960.96083832],
             [1118698.5715939 ],
             [2082407.37139755],
             [1026686.38365648],
             [1153551.2773749 ],
             [1568470.1086974 ],
             [1076947.13448431],
             [1330879.82644885],
             [1750747.42422742],
             [1014812.62871642],
             [1801926.55600264],
             [1620939.46498849],
             [1423668.23622426],
             [1108698.55612707],
             [1541929.98626904],
```



[47]:
```
from sklearn.model_selection import cross_val_score
```

[48]:
```
cross_val_value_training=cross_val_score(linearmodel,X_train,y_train,cv=10)
```

[49]:
```
cross_val_value_training.mean()
```

[49]: 0.9182174480513696

[50]:
```
cross_val_value_testing=cross_val_score(linearmodel,X_test,y_test,cv=10)
```

[51]:
```
cross_val_value_testing.mean()
```

[51]: 0.9114034511775113

[12]:
```
X=pd.read_csv('../input/usa-housing/USA_Housing.csv')
```

[13]:
```
X=X.drop("Address",axis=1)
```

[14]:
```
X.head()
```

[14]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
0	79545.458574	5.682861	7.009188	4.09	23086.800503	1.059034e+06
1	79248.642455	6.002900	6.730821	3.09	40173.072174	1.505891e+06
2	61287.067179	5.865890	8.512727	5.13	36882.159400	1.058988e+06
3	63345.240046	7.188236	5.586729	3.26	34310.242831	1.260617e+06
4	59982.197226	5.040555	7.839388	4.23	26354.109472	6.309435e+05

[15]:
```
X=X.drop("Price",axis=1)
```

[16]:
```
X.head()
```

[16]:

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population
0	79545.458574	5.682861	7.009188	4.09	23086.800503
1	79248.642455	6.002900	6.730821	3.09	40173.072174
2	61287.067179	5.865890	8.512727	5.13	36882.159400
3	63345.240046	7.188236	5.586729	3.26	34310.242831
4	59982.197226	5.040555	7.839388	4.23	26354.109472

[17]:
```
y=houses[["Price"]]
```

[18]:
```
y.head()
```

[18]:

	Price
0	1.059034e+06
1	1.505891e+06
2	1.058988e+06
3	1.260617e+06

CONCLUSION:

In conclusion, house price prediction is a crucial aspect of the real estate market that can benefit both buyers and sellers. By leveraging advanced machine learning and data analysis techniques, we can make more accurate and informed decisions when it comes to buying or selling a home. Predictive models, such as regression models, neural networks, or ensemble methods, can help us estimate property values based on a wide range of factors, including location, size, amenities, and market trends