PREDICTING HOUSE PRICE USING MACHINE LEARNING

PROJECT : HOUSE PRICE PREDICTION

OBJECTIVE:

The objective of the project is to develop a machine learning moded that Accurately protect the prices of house based on house set of features such as location, Square footage number of bedroom and bathroom and other revalent factor.

Phase 1: DATA PREPROCESSING AND FEATURE ENGINEERING

House Price Prediction using Machine Learning

So to deal with this kind of issues Today we will be preparing a MACHINE LEARNING Based model, trained on the House Price Prediction Dataset.

1	Id	To count the records.
2	MSSubClass	Identifies the type of dwelling involved in the sale.
3	MSZoning	Identifies the general zoning classification of the sale.
4	LotArea	Lot size in square feet.
5	LotConfig	Configuration of the lot
6	BldgType	Type of dwelling
7	OverallCond	Rates the overall condition of the house
8	YearBuilt	Original construction year
9	YearRemodAdd	Remodel date (same as construction date if no remodeling or additions).
10	Exterior1st	Exterior covering on house
11	BsmtFinSF2	Type 2 finished square feet.
12	TotalBsmtSF	Total square feet of basement area
13	SalePrice	To be predicted

Importing Libraries and Dataset

Here we are using

- **Pandas** – To load the Dataframe
- **Matplotlib** – To visualize the data features i.e. barplot
- **Seaborn** – To see the correlation between features using heatmap
    - Python3

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

dataset = pd.read_excel("HousePricePrediction.xlsx")

# Printing first 5 records of the dataset
print(dataset.head(5))
```

OUTPUT:

	MSSubClass	MSZoning	LotArea	LotConfig	BldgType	OverallCond	YearBuilt
0	60	RL	8450	Inside	1Fam	5	2003
1	20	RL	9600	FR2	1Fam	8	1976
2	60	RL	11250	Inside	1Fam	5	2001
3	70	RL	9550	Corner	1Fam	5	1915
4	60	RL	14260	FR2	1Fam	5	2000

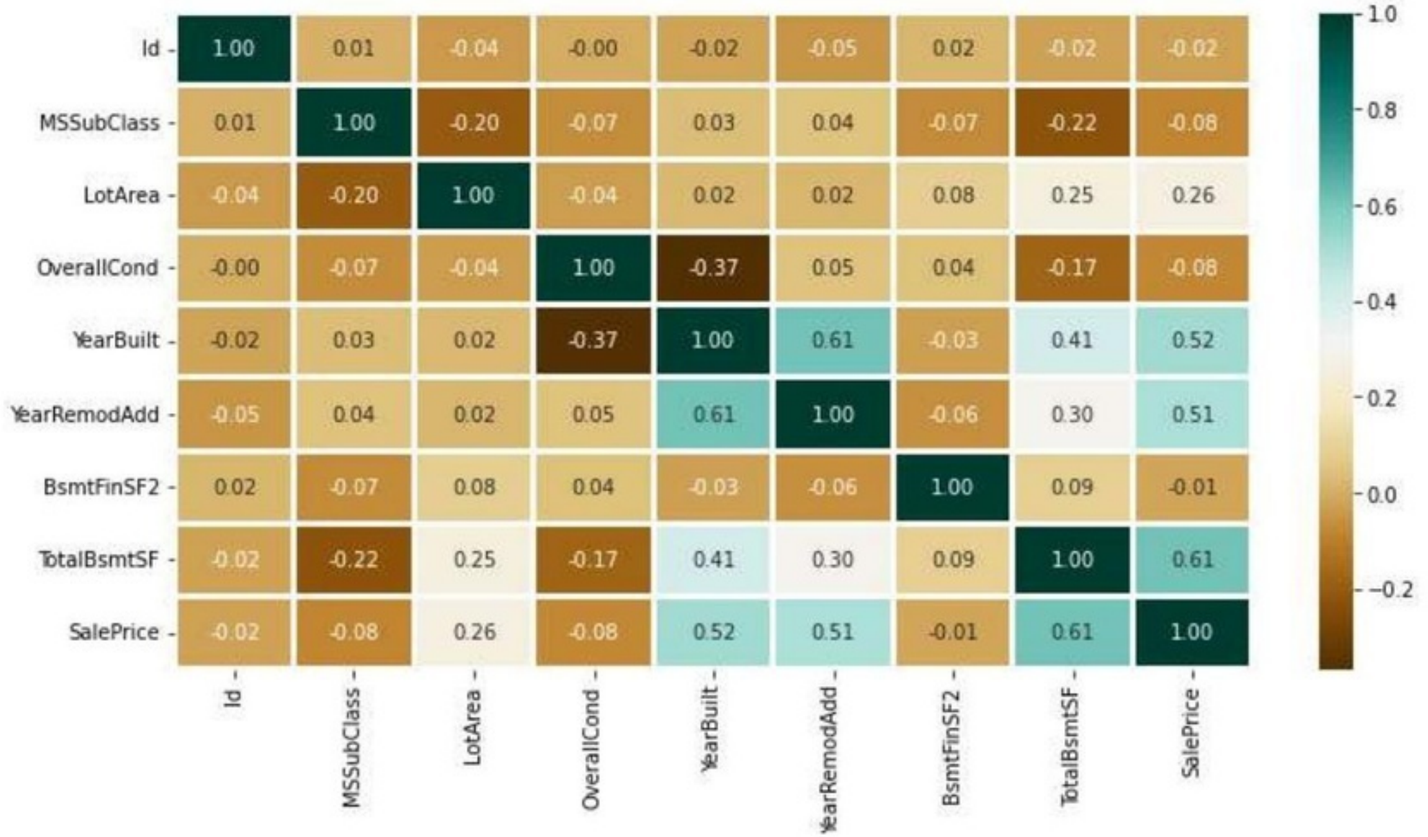
	YearRemodAdd	Exterior1st	BsmtFinSF2	TotalBsmtSF	SalePrice
0	2003	VinylSd	0.0	856.0	208500.0
1	1976	MetalSd	0.0	1262.0	181500.0
2	2002	VinylSd	0.0	920.0	223500.0
3	1970	Wd Sdng	0.0	756.0	140000.0
4	2000	VinylSd	0.0	1145.0	250000.0

Data Preprocessing

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

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

OUTPUT:


Id
MSSubClass
LotArea
OverallCond
YearBuilt
YearRemodAdd
BsmtFinSF2
TotalBsmtSF
SalePrice
1.00 0.01 -0.04 -0.00 -0.02 -0.05 0.02 -0.02 -0.02
0.01 1.00 -0.20 -0.07 0.03 0.04 -0.07 -0.22 -0.08
-0.04 -0.20 1.00 -0.04 0.02 0.02 0.08 0.25 0.26
-0.00 -0.07 -0.04 1.00 -0.37 0.05 0.04 -0.17 -0.08
-0.02 0.03 0.02 -0.37 1.00 0.61 -0.03 0.41 0.52
-0.05 0.04 0.02 0.05 0.61 1.00 -0.06 0.30 0.51
0.02 -0.07 0.08 0.04 -0.03 -0.06 1.00 0.09 -0.01
-0.02 -0.22 0.25 -0.17 0.41 0.30 0.09 1.00 0.61
-0.02 -0.08 0.26 -0.08 0.52 0.51 -0.01 0.61 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Conclusion

Clearly, SVM model is giving better accuracy as the mean absolute error is the least among all the other regressor models i.e. 0.18 approx. To get much better results ensemble learning techniques like Bagging and Boosting can also be used.