Project Report

ON

Mumbai House Price Prediction

DEV BHOOMI INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering

(Batch 2020-24)

Submitted By :

Shashi Ranjan (Roll : 200080101078)

Sandeep Yadav (Roll : 200080101070)

Sumitesh Raj (Roll : 200080101087)

Under The Guidance Of:

Mr. Subhashish Goswami

Associate Professor

CONTENT

ACKNOWLEDGEMENT

We deem it a pleasure to acknowledge our sense of gratitude to our project guide Mr. Subhashish Goswami under whom we have carried out the project work. His incisive and objective guidance and timely advice encouraged us with constant flow of energy to continue the work.

We wish to reciprocate in full measure the kindness shown by Dr. Ritika Mehra (DEAN, Computer Science and Engineering) who inspired us with his valuable suggestions in successfully completing the project work. We shall remain grateful to Dev Bhoomi Institute Of Technology, for providing us a strong academic atmosphere by enforcing strict discipline to do the project work with utmost concentration and dedication.

Finally, we must say that no height is ever achieved without some sacrifices made at some end and it is here where we owe our special debt to our parents and our friends for showing their generous love and care throughout the entire period of time.

Date:

Shashi Ranjan

Roll : 200080101078

Sumitesh Raj

Roll : 200080101087

Sandeep Yadav

Roll : 200080101070

ABSTRACT

House Price Index is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from House price prediction to predict individual housing price. There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern about the performance of individual models and neglect the less popular yet complex models. As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

INTODUCTION

House price prediction is great project to learn and apply the machine learning algorithm. The basic idea behind this project is we are training the machine using the machine learning algorithm from the data set.

In this busy world it is very difficult to find a house according to our need and budget. It becomes more difficult to find the house in metropolitan cities like Mumbai, Kolkata, Delhi, etc. This project uses the data of Mumbai city in order to train and test the machine so that it become capable of predicting the price of house. Machine learning algorithm makes it easy to know the price of houses depending on the location, area, number of bedrooms, etc.

In this project Random Forest Regression, Linear Regression, and Decision Tree Machine learning algorithm has been used to compare the efficiency of the algorithm. Based on comparison we predict which algorithm best suits for the prediction of price of house in Mumbai.

LITERATURE WORK

Pattern recognition is the process of recognizing patterns by using a machine learning algorithm. Pattern recognition can be defined as the classification of data based on knowledge already gained or on statistical information extracted from patterns and/or their representation. One of the important aspects of pattern recognition is its application potential.

In Pattern Recognition, pattern is comprises of the following two fundamental things:

- Collection of observations
- The concept behind the observation

Features may be represented as continuous, discrete, or discrete binary variables. A feature is a function of one or more measurements, computed so that it quantifies some significant characteristics of the object.

Design Principles of Pattern Recognition

In pattern recognition system, for recognizing the pattern or structure two basic approaches are used which can be implemented in different techniques. These are

I. Statistical Approach

II. Structural Approach

Statistical Approach:

Statistical methods are mathematical formulas, models, and techniques that are used in the statistical analysis of raw research data. The application of statistical methods extracts information from research data and provides different ways to assess the robustness of research outputs.

Two main statistical methods are used :

Descriptive Statistics: It summarizes data from a sample using indexes such as the mean or standard deviation.

Inferential Statistics: It draw conclusions from data that are subject to random variation.

Structural Approach:

The Structural Approach is a technique wherein the learner masters the pattern of sentence. Structures are the different arrangements of words in one accepted style or the other.

Types of structures:

Sentence Patterns

Phrase Patterns

Formulas

Idioms

IMPLEMENTATION SET UP

Dataset

Dataset is a collection of related sets of information which is used for the purpose of training and testing the machine learning algorithm. Based on this some prediction is made

List of Attributes

Attribute Name	Data Type	Description
Price	int64	Price of the house
Area	int64	Area of house
Location	object	Location of house in Mumbai
No. of Bedrooms	int64	Total number of bedrooms in the house
New/Resale	int64	Whether the house is new of resale
Gymnasium	int64	Availability of nearby Gymnasium
Lift Available	int64	Availability of lift in the Building
Car Parking	int64	Availability of space for car parking
Maintenance Staff	int64	Availability of maintenance staff
24 X 7 Security	int64	24x7 Security enabled
Clubhouse	int64	Nearby clubhouse availability
Intercom	int64	Nearby clubhouse availability
Landscaped Gardens	int64	Availability of Gardens
Indoor Games	int64	Indoor game house
Gas Connection	int64	Availability of gas connection
Jogging Track	int64	Nearby jogging track available
Swimming Pool	int64	Building having Swimming Pool

The dataset used in the project is taken from github. The dataset consists of 6338 row and 17 columns. All the attribute, data type and description of attribute is listed above.

.Dataset is broken into two parts :

(1) Training dataset : Training data set is used to train the model to predict something out of it.

(2) Testing dataset : Testing data set is used to test the model to designed and trained using the training dataset to predict the output.

Generally dataset is split into training and testing in 80:20, 70:30 ratio.

Parameters

Parameters are the attributes present in the dataset. Generally attributes present is called parameters but attribute supplied by user is also an parameter called Hypermeter. For example, value of ʻK' in KNN algorithm.

Parameters use in the dataset can be classified into two types :

1. Categorical variables : A categorical variable refers to characteristic that can't be quantifiable. Categorical variables can Binary, Nominal or Ordinal. This is often called as qualitative variable.

I. Binary Variables : This is a categorical variable that takes only two value i .e., either 1 or 0. 1- represent True and 0 - represent False.

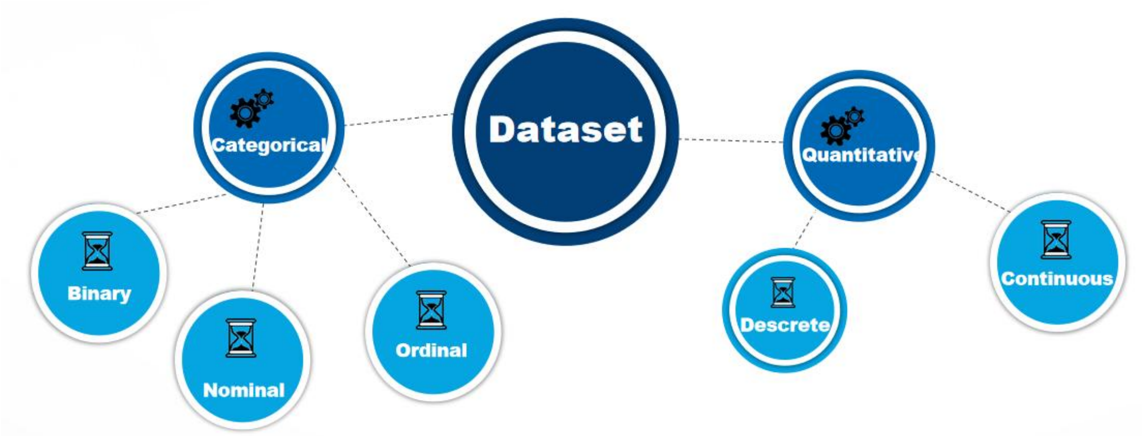
Binary variables present in the dataset are: New/Resale, Gymnasium, Car Parking, Maintenance, 24x7 Security, Club House, Intercom, Landscaped, Indoor Games, Gas Connection, Jogging Track, Swimming Pool.

II. Nominal Variables : A nominal variable is one that describes a name, label or category without natural order.

Only ʻLocation' is the nominal variable in the dataset used.

III. Ordinal Variables : A ordinal variable is whose values are defined by an order relation between the different categories.

There is no any Ordinal data present in the dataset.



2. Quantitative Variables : A quantitative variable is quantifiable characteristics whose values are number. Numeric variables may either be Discrete or Continuous. It is often called as Numeric Variables.

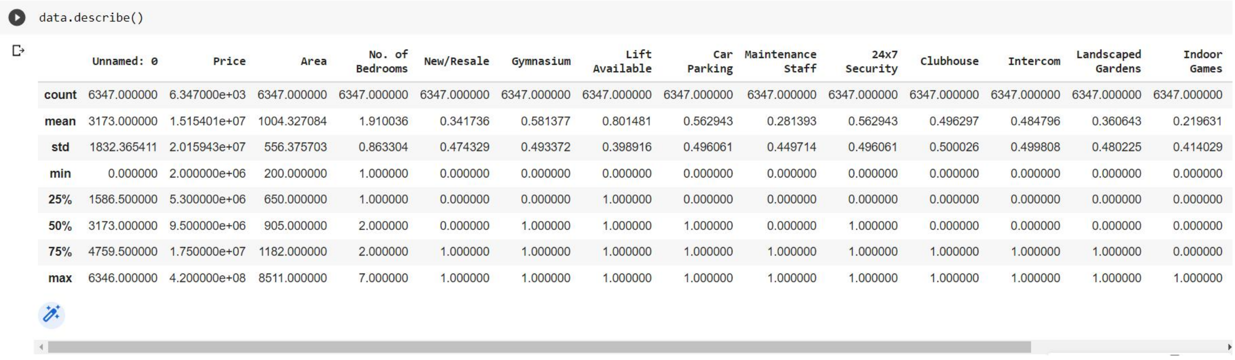
I. Continuous Variables : A variable is said to be continuous if it can assume an infinite number of real values within a given interval.

Continuous variables present in dataset are: Price and Area

II. Discrete Variables : A discrete variable can assume only a finite number of real values within a given in interval.

Continuous variables present in dataset are: No. of Bedrooms and Lift Available

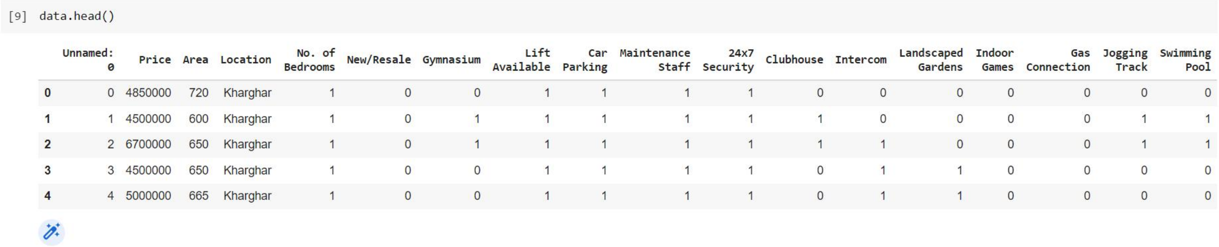
Data description in shown below:



```
data.describe()
```

	Unnamed: 0	Price	Area	No. of Bedrooms	New/Resale	Gymnasium	Lift Available	Car Parking	Maintenance Staff	24x7 Security	Clubhouse	Intercom	Landscaped Gardens	Indoor Games
count	6347.000000	6.347000e+03	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000	6347.000000
mean	3173.000000	1.515401e+07	1004.327084	1.910036	0.341736	0.581377	0.801481	0.562943	0.281393	0.562943	0.496297	0.484796	0.360643	0.219631
std	1832.365411	2.015943e+07	556.375703	0.863304	0.474329	0.493372	0.398916	0.496061	0.449714	0.496061	0.500026	0.499808	0.480225	0.414029
min	0.000000	2.000000e+06	200.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1586.500000	5.300000e+06	650.000000	1.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	3173.000000	9.500000e+06	905.000000	2.000000	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
75%	4759.500000	1.750000e+07	1182.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
max	6346.000000	4.200000e+08	8511.000000	7.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

In the data description image number of rows present corresponding to each attribute is shown by count column. Mean depict mean value of each attributes, min and max depicts minimum and maximum value respectively corresponding to each row. 25%, 50% and 75% shows what values 25% , 50%, and 75% of data contains in the row corresponding to each attributes. Below is the image showing head part of data.



```
[9] data.head()
```

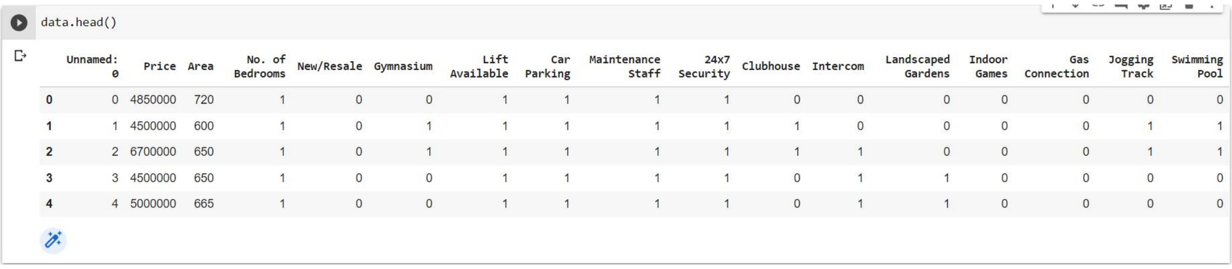
	Unnamed: 0	Price	Area	Location	No. of Bedrooms	New/Resale	Gymnasium	Lift Available	Car Parking	Maintenance Staff	24x7 Security	Clubhouse	Intercom	Landscaped Gardens	Indoor Games	Gas Connection	Jogging Track	Swimming Pool
0	0	4850000	720	Kharghar	1	0	0	1	1	1	1	0	0	0	0	0	0	0
1	1	4500000	600	Kharghar	1	0	1	1	1	1	1	1	0	0	0	0	1	1
2	2	6700000	650	Kharghar	1	0	1	1	1	1	1	1	1	0	0	0	1	1
3	3	4500000	650	Kharghar	1	0	0	1	1	1	1	0	1	1	0	0	0	0
4	4	5000000	665	Kharghar	1	0	0	1	1	1	1	0	1	1	0	0	0	0

Data Preprocessing

Data Preprocessing includes the steps we need to follow to transform or encode data so that it may be easily parsed by the machine.

The main agenda for a model to be accurate and precise in predictions is that the algorithm should be able to easily interpret the features of the data.

This project to need to preprocess the data in order to apply the machine learning algorithm. The dataset used contains an attribute called **'Location'** needs to be either removed. Dataset after preprocessing is shown in image.



```
data.head()
```

	Unnamed: 0	Price	Area	No. of Bedrooms	New/Resale	Gymnasium	Lift Available	Car Parking	Maintenance Staff	24x7 Security	Clubhouse	Intercom	Landscaped Gardens	Indoor Games	Gas Connection	Jogging Track	Swimming Pool
0	0	4850000	720	1	0	0	1	1	1	1	0	0	0	0	0	0	0
1	1	4500000	600	1	0	1	1	1	1	1	1	0	0	0	0	1	1
2	2	6700000	650	1	0	1	1	1	1	1	1	1	0	0	0	1	1
3	3	4500000	650	1	0	0	1	1	1	1	0	1	1	0	0	0	0
4	4	5000000	665	1	0	0	1	1	1	1	0	1	1	0	0	0	0

Data Visualization

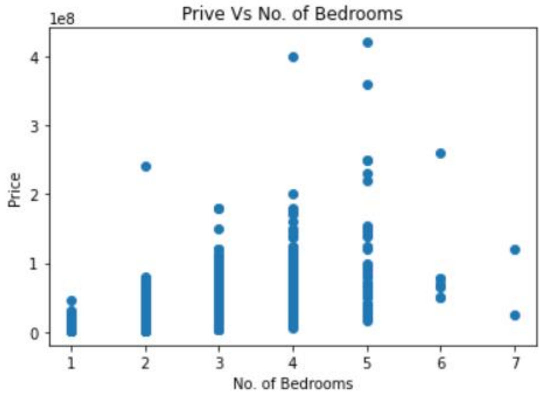
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

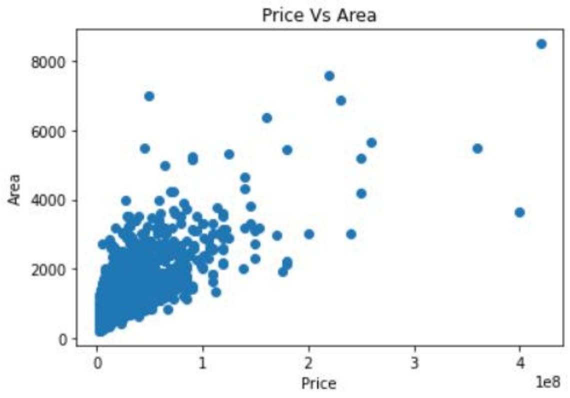
Some of the visualization of the the dataset used is shown below :

Below two images shows the *scatter plot* for the attributes of the dataset.

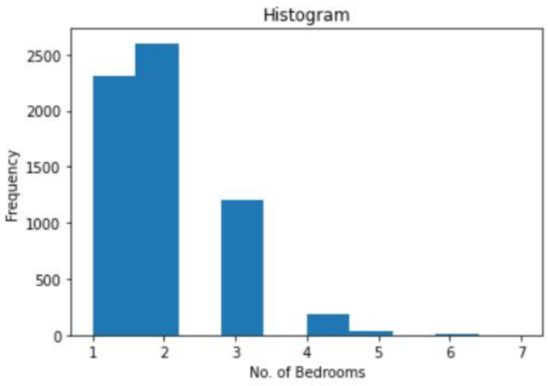
1. *Price Vs No. of Bedrooms* shows that the price of houses increases with increase in number of bedroom. At the same time it also shows that some of the houses having 6 and 7 bedrooms still have lesser price, it is due to the locality in which the house is present.

2. *Price Vs Area* shows that the price of houses increases with increase in the area of house. It shows that there is good correlation in the attribute of the dataset Price Vs Area. This data also has some outliers.

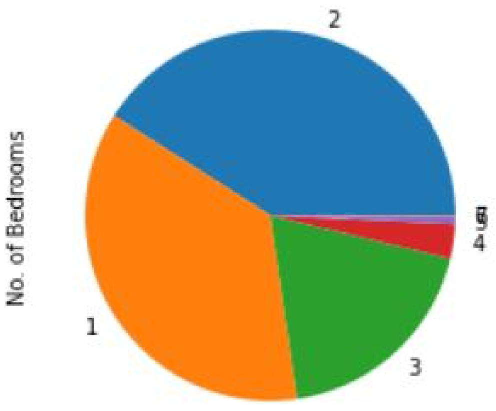




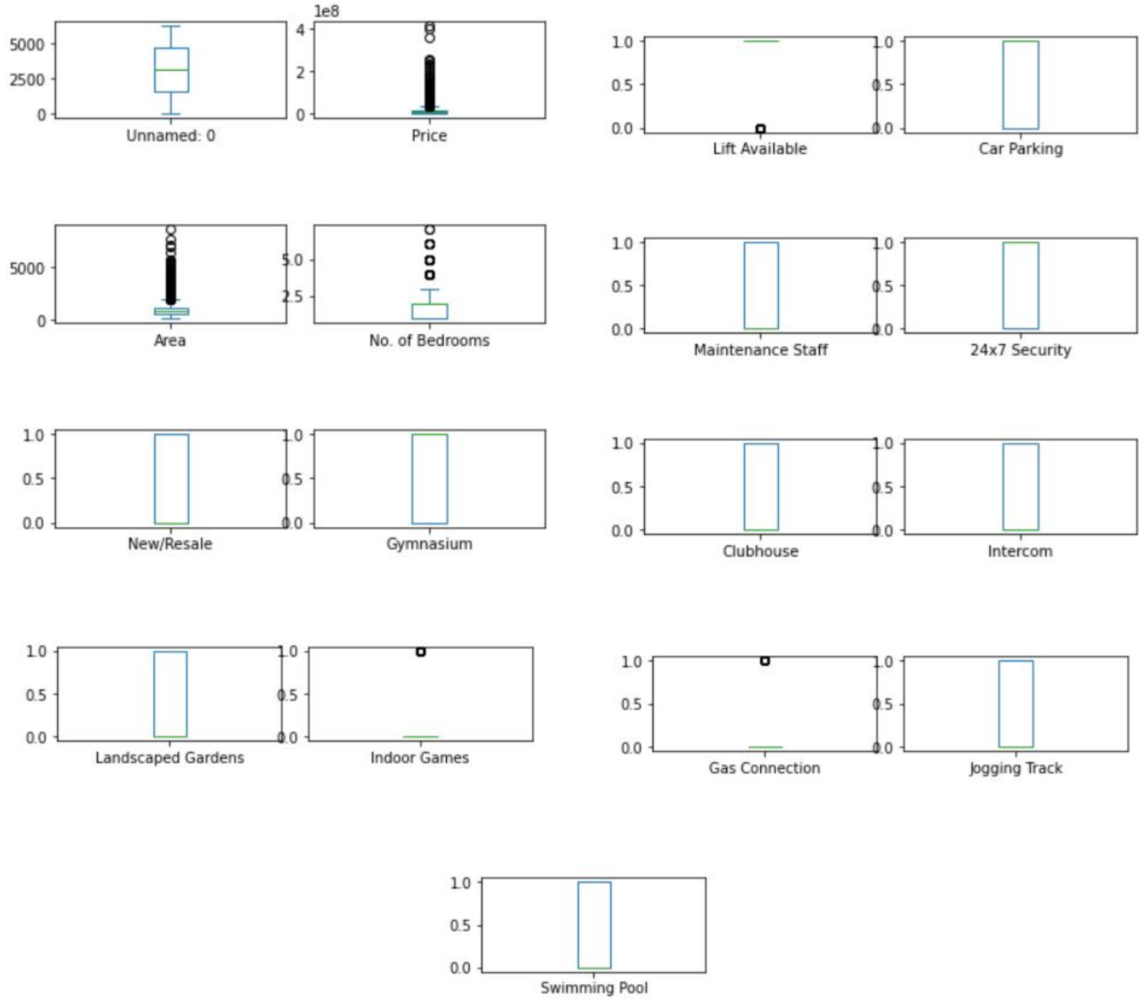
Histogram shows the number of houses available with 1,2,3,4,5,6,and 7 bedrooms. It can be seen that most of the houses have 1 or 2 bedrooms, there is very few houses having 5,6,7 bedrooms



Pie chart also shows the number of houses with available number of bedrooms..



Box and Wisher Plot for all the attributes are shown below. The plot shows Upper extreme, Upper Quartile, Median, Lower Quartile, Lower Extreme and Outliers.



Steps to apply the Machine Learning Algorithm :

Step 1. To apply the machine learning algorithms we need to select a independent (x) and a target variable (y).

```
X = data.iloc[:,1:].values

y = data.iloc[:,0].values
```

Step 2. Now we need to split the dataset into two parts that is training and testing.Training dataset will be used to train the model and testing will be uset to test the result for the model. Code for splitting the dataset to test and train data :

```
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,random_s
tate=0)
```

Step 3. Apply the Machine leaning algorithms .

In this project the output of the model is tested by applying three different models : (1) Multiple Linear Regression, (2) Random Forest, and (3) Decision Tree.

Different Machine learning algorithms used are discussed briefly :

1. Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. It is based on supervised machine learning algorithm. These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula :

y = c + b*x

where,

y = estimated dependent variable score,

c = constant,

b = regression coefficient,

x = score on the independent variable.

In the project *Multiple linear regression* algorithm has been used .Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Code for implementing Multiple Linear Regression :

```
mlr = LinearRegression()
mlr.fit(X_train,y_train)
mlr_score = mlr.score(X_test,y_test)
pred_mlr = mlr.predict(X_test)
expl_mlr = explained_variance_score(pred_mlr,y_test)
```

2. Decision Tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

Important Terms Used in Decision Trees

1. Entropy: Entropy is the measure of uncertainty or randomness in a data set. Entropy handles how a decision tree splits the data.

It is calculated using the following formula:

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

Where Entropy is denoted by H(S) for a finite set S

2. Information Gain: The information gain measures the decrease in entropy after the data set is split.

It is calculated as follows:

$$IG(S,A) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$$

where ,

IG(S, A) is the information gain by applying feature A

H(S) is the Entropy of the entire set,

P(x) is the probability of event x

3. Gini Index: The Gini Index is used to determine the correct variable for splitting nodes. It measures how often a randomly chosen variable would be incorrectly identified.

4. Root Node: The root node is always the top node of a decision tree. It represents the entire population or data sample, and it can be further divided into different sets.

5. Decision Node: Decision nodes are subnodes that can be split into different subnodes; they contain at least two branches.

6. Leaf Node: A leaf node in a decision tree carries the final results. These nodes, which are also known as terminal nodes, cannot be split any further.

Code for implementing Decision Tree:

```
tr_regressor = DecisionTreeRegressor(random_state=0)
tr_regressor.fit(X_train,y_train)
tr_regressor.score(X_test,y_test)
pred_tr = tr_regressor.predict(X_test)
decision_score=tr_regressor.score(X_test,y_test)
expl_tr = explained_variance_score(pred_tr,y_test)
```

3. Random Forest

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability. Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

Code for implementing Random Forest :

```
rf_regressor = RandomForestRegressor(n_estimators=28,random_state=0)
rf_regressor.fit(X_train,y_train)
rf_regressor.score(X_test,y_test)
rf_pred =rf_regressor.predict(X_test)
rf_score=rf_regressor.score(X_test,y_test)
expl_rf = explained_variance_score(rf_pred,y_test)
```

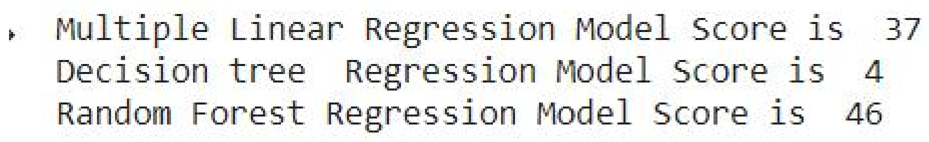
Step 4. After applying all the Machine Learning Algorithms we need to compare the results obtained from different models to make certain conclusion. Result of the model score is compared.

RESULT ANALYSIS

We have applied the three machine learning algorithms successfully and the results of these algorithms can be observed in Model Score. Model score for Multiple Linear Regression is 37, for Decision Tree it is 4 which is very low, and Model score for Random Forest is 46 which is highest. Let's have a look to the model score through code:

```
print("Multiple Linear Regression Model Score is ",round(mlr.score(X_test,
y_test)*100))

print("Decision tree  Regression Model Score is ",round(tr_regressor.scor
e(X_test,y_test)*100))

print("Random Forest Regression Model Score is ",round(rf_regressor.score
(X_test,y_test)*100))
```

Output for the above code :

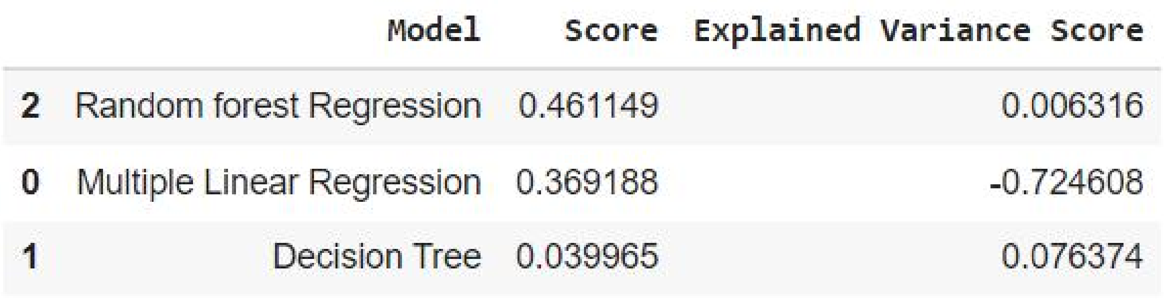


```
Multiple Linear Regression Model Score is  37
Decision tree  Regression Model Score is  4
Random Forest Regression Model Score is  46
```

Code for clear comparison among different models used over the same datset :

```
models_score = pd.DataFrame({

'Model':['Multiple Linear Regression','Decision Tree',
'Random forest Regression'], 'Score':[mlr_score,decision_score,rf_score],

'Explained Variance Score':[expl_mlr,expl_tr,expl_rf]

})

models_score.sort_values(by='Score',ascending=False)
```

Output for above code in tabular format :



	Model	Score	Explained Variance Score
2	Random forest Regression	0.461149	0.006316
0	Multiple Linear Regression	0.369188	-0.724608
1	Decision Tree	0.039965	0.076374

From the above comparisons we can say that best fit algorithm for the dataset taken is Random Forest with Model Score of 46%.

CONCLUSION AND FUTURE SCOPE

The model designed accuracy depends on the dataset selected, better the dataset better will be the accuracy. Best suited model applied is Random Forest. This can be applied to datset of any city for their house price prediction. The project can be enhanced by UI designing through they can predict the price in more easier and interactive way. In this busy world it will be of immense use to search for a house at near to our workplace.

REFERENCES

https://github.com/

https://colab.research.google.com/

https://www.simplilearn.com/

https://www.kaggle.com/

https://www.sciencedirect.com/

https://www.geeksforgeeks.org/