**MACHINE LEARNING**

**DIAMONDS PRICE PREDICTION**

**2020**

**SUBMITTED BY:-**

1. **SAHIL MANGLA – 18104101**
2. **HARSHIT GARG – 18104102**
3. **GAUTAM SHARMA --18104104**
4. **ARJUN VIJ – 18104122**

5/15/2020

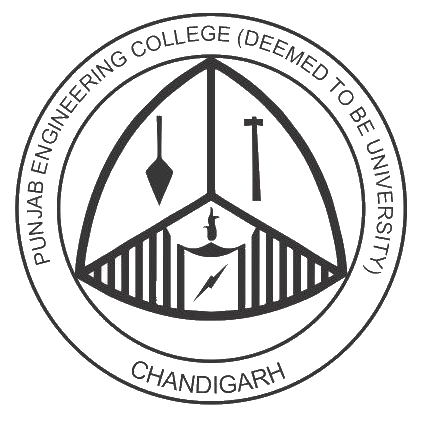
**MODEL**



**2ND**

**YEAR**

**ELECTRICAL**



**PUNJAB**

**ENGINEERING**

**COLLEGE, CHANDIGARH**

**SUBMITTED TO:-**

MRS. PADMAVATI KHADNOOR **DIAMONDS PRICE PREDICTION**

**MACHINE LEARNING MODEL**

**REPORT**

1. ***INTRODUCTION* :-**

**Diamond** is a [solid form of the element carbon](https://en.wikipedia.org/wiki/Allotropes_of_carbon) with its atoms arranged in a [crystal structure](https://en.wikipedia.org/wiki/Crystal_structure) called [diamond cubic](https://en.wikipedia.org/wiki/Diamond_cubic). At [room temperature and pressure](https://en.wikipedia.org/wiki/Standard_conditions_for_temperature_and_pressure), another solid form of carbon known as [graphite](https://en.wikipedia.org/wiki/Graphite) is the [chemically stable](https://en.wikipedia.org/wiki/Chemical_stability) form, but diamond almost never converts to it. Diamond has the highest [hardness](https://en.wikipedia.org/wiki/Scratch_hardness) and [thermal conductivity](https://en.wikipedia.org/wiki/Thermal_conductivity) of any natural material, properties that are utilized in major industrial application such as cutting and polishing tools.

Diamond is a stone which is very popular among the people for different purposes especially as jewelry. There are different categories of diamonds which makes it either cheaper or very expensive Major industrial applications such as cutting and polishing tools.

The aims of the project are:

1. To identify the best model by equating the best accuracy for different regression model.
2. To predict the price of diamond by the best model from different features.
3. To do data preprocessing in the dataset to make it better (Data aggregation, changing column names and changing the class categories as per the requirement)
4. Visualization of the data through graphical representation to identify the relationship between price and carat with respect to color, cut and clarity.
5. ***PROBLEM STATEMENT:-***

Data set of diamond has been taken in this final project to identify the relationship between the price of diamonds with other factors such as carat, color, cut, clarity, x, y, z and depth. This project focuses on identifying factors causing changes in the price of diamonds in the current data set which is downloaded from data source online.

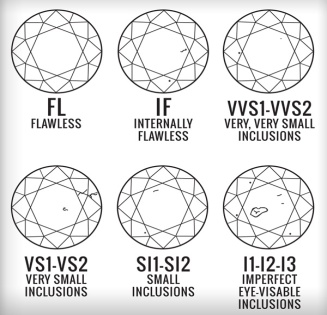
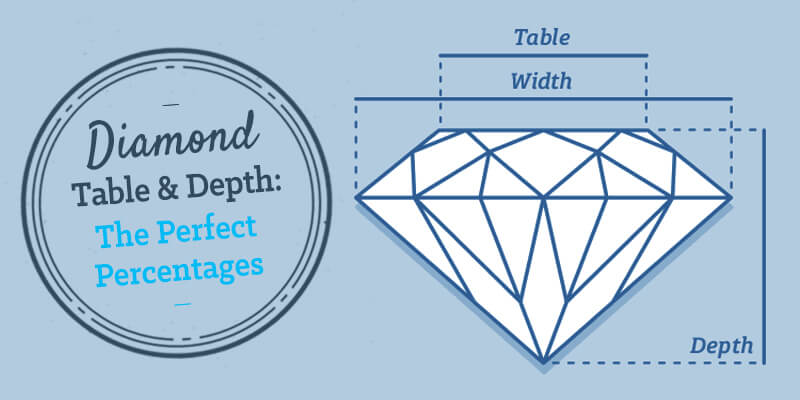
1. ***DATASET USED:-***
2. DATASET LINK:-

[***https://www.kaggle.com/shivam2503/diamonds***](https://www.kaggle.com/shivam2503/diamonds)

1. DATA DESCRIPTION:-

A data frame with 53940 rows and 10 variables. Having the different columns in the dataset are following:-

1. price (in US dollars (\$326--\$18,823))



1. carat (weight of the diamond (0.2--5.01))
2. cut (quality of the cut (Fair, Good, Very Good, Premium, Ideal))
3. color (diamond color, from J (worst) to D (best))
4. clarity( a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)))
5. x (length in mm (0--10.74))
6. y (width in mm (0--58.9))
7. z (depth in mm (0--31.8))
8. depth (total depth percentage = z / mean(x, y) = 2 \* z / (x + y))\
9. table width of top of diamond relative to widest point (43--95)
10. ***METHODOLOGY:-***
11. **DATA PREPROCESSING:-**

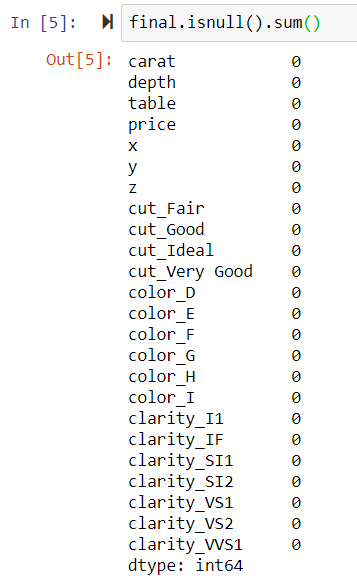
Firstly we have to preprocess the data so that we can easily train our data to different regression algorithms

In data preprocessing we have different steps to be followed :-

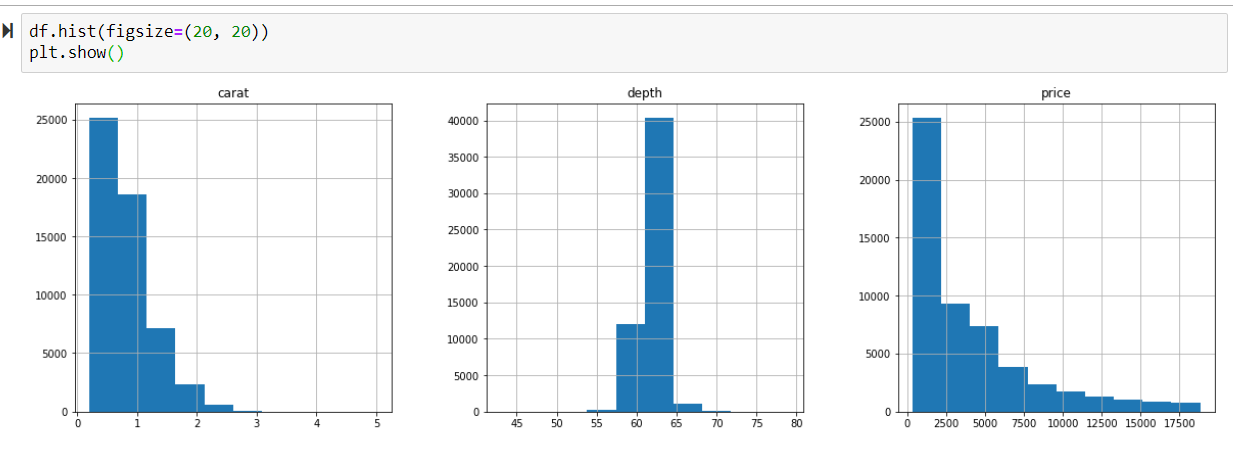
1. ***First step:-*** We have to see if there is any missing data in the dataset or not.

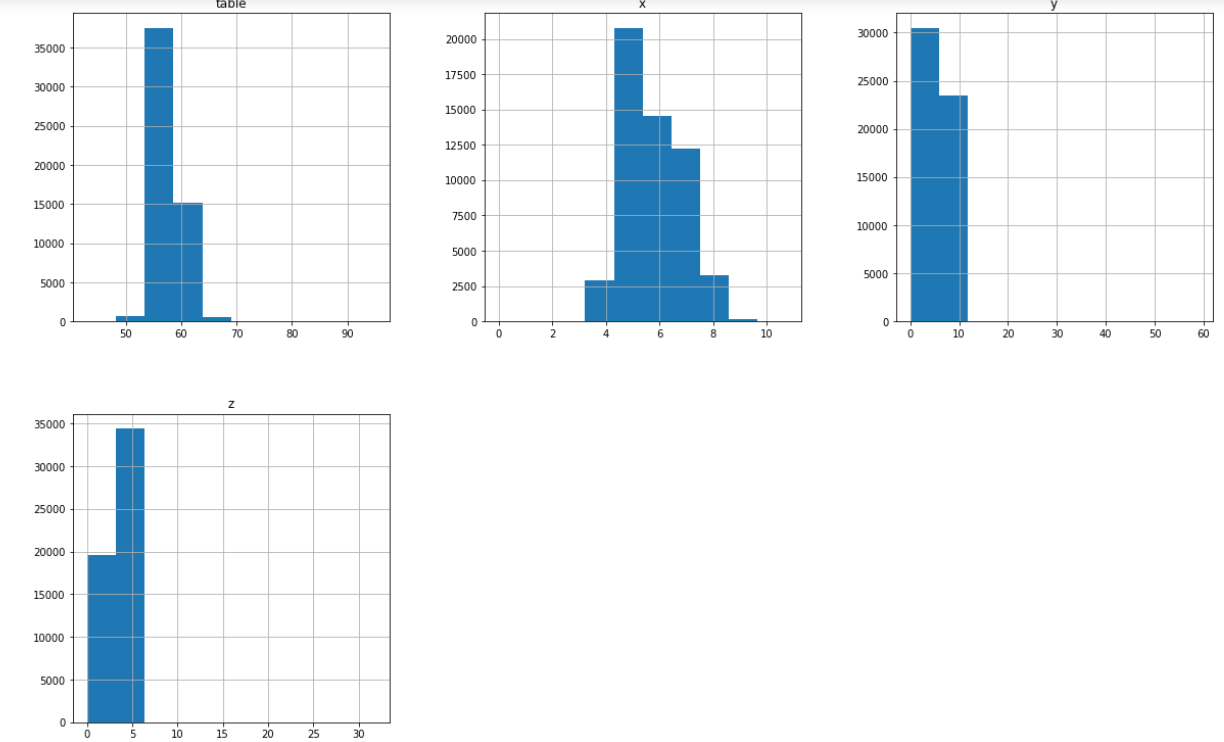
So to find this there are two methods to find out

* First method- we can use the pandas functions isnull and sum to calculate the total the missing values in different columns.

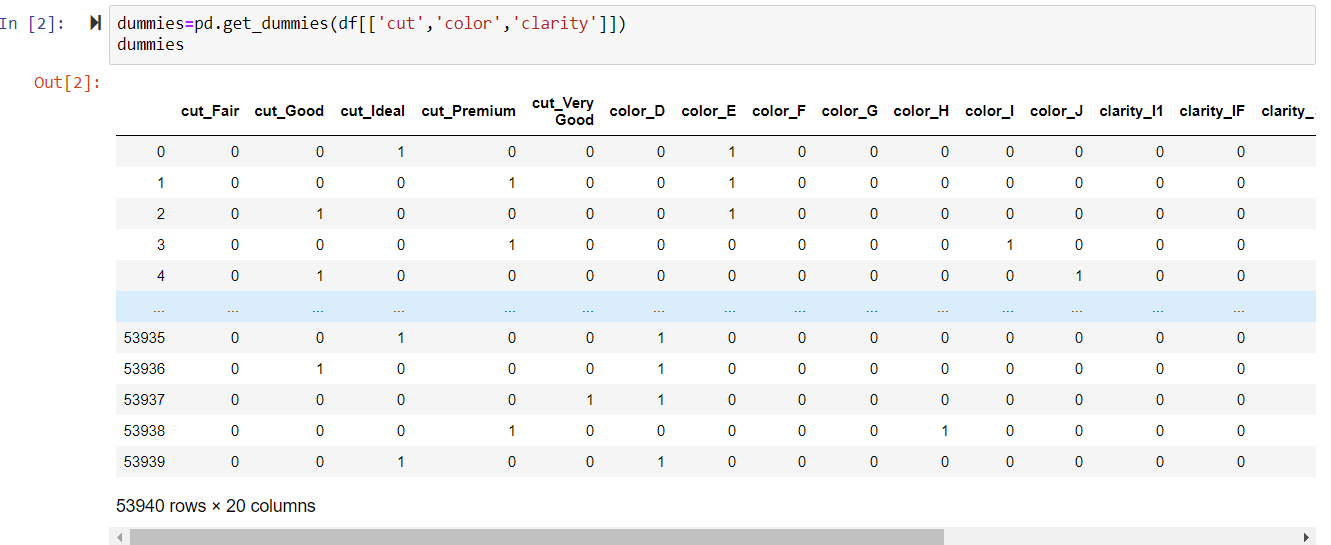


* Second method- we can draw different histograms from matplotlib library to find is there any missing value or we can use countplot() function in seaborn library to find out missing value

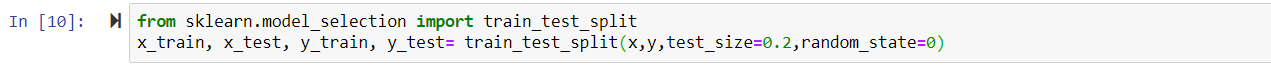




1. ***Second step:-*** we have to solve the problem of categorical data. The method to solve the problem of categorical data is the function of module pandas i.e dummyvariable(). By which we can create dummies of the categorical data. And make the data ready to train the model.

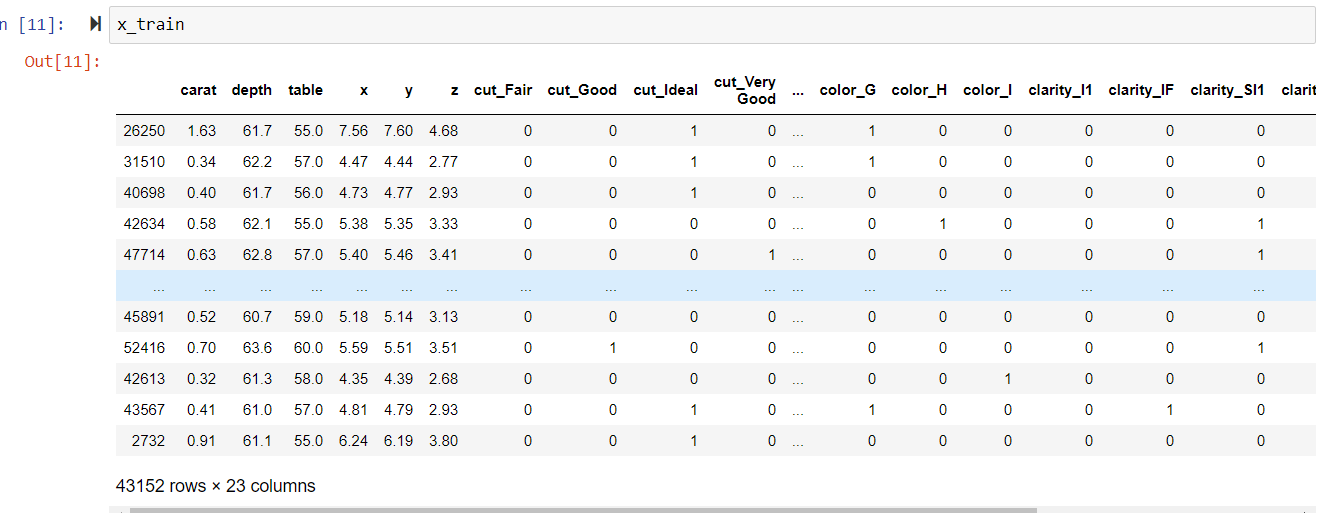


1. ***Third step:-*** is to normalize data but in the dataset there is no need of normalizing the data as the data is not varying by the large value.
2. ***Fourth step:-*** is to divide the dataset into two part by using the function train\_test\_split() of sklearn.preprocessing module .



Test size=20% of the total rows in the dataset

So now the data is ready so that it can be train and test on different algorithms of machine learning



1. **DATA VISUALIZATION:-**

Before training our model to some algorithms we can use the data visualization techniques to observe the dataset by graphical.

*WHY DATA VISUALIZATION?*

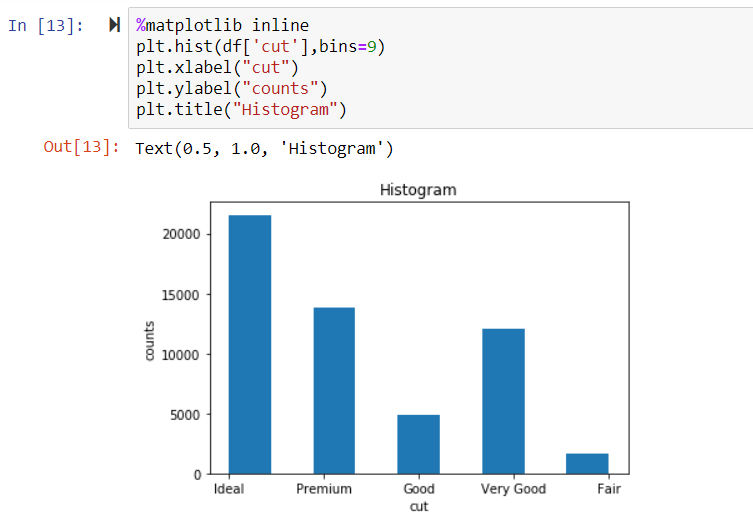
Data visualization allows us to quickly interpret the data and adjust variables to see their effect.

*PLOTTING:-*

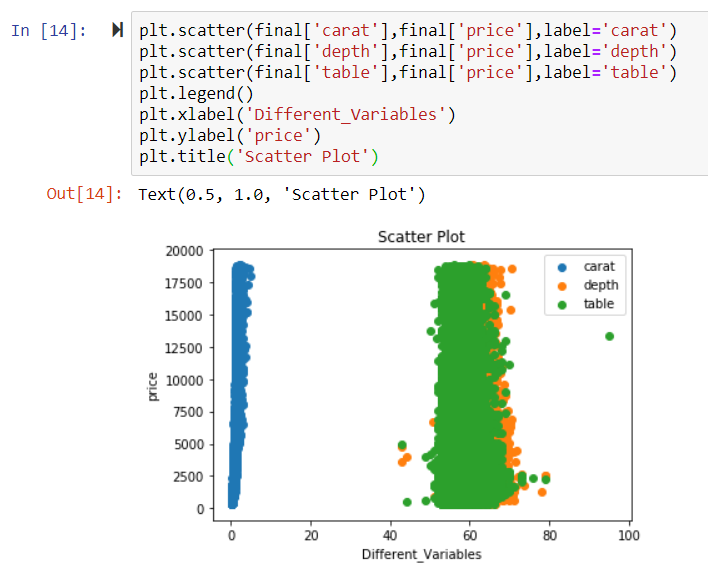
Data can be visualized by the two libraries mainly named matplotlib and seaborn.

There are different graphs we can use from matplotlib library:-

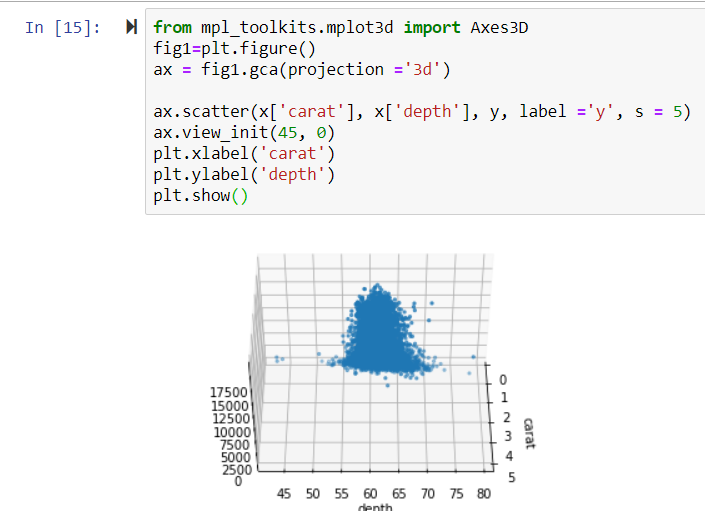
1. Histogram



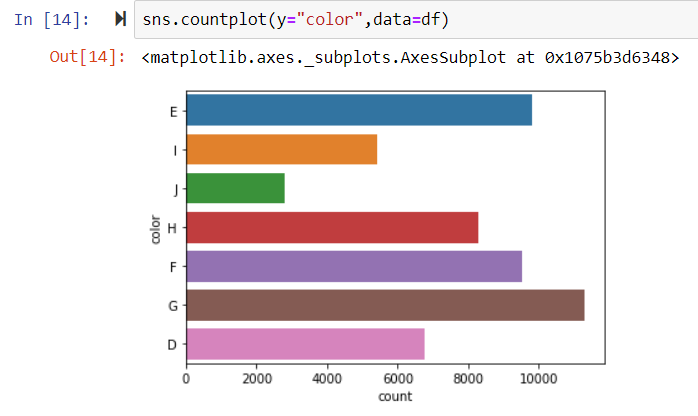
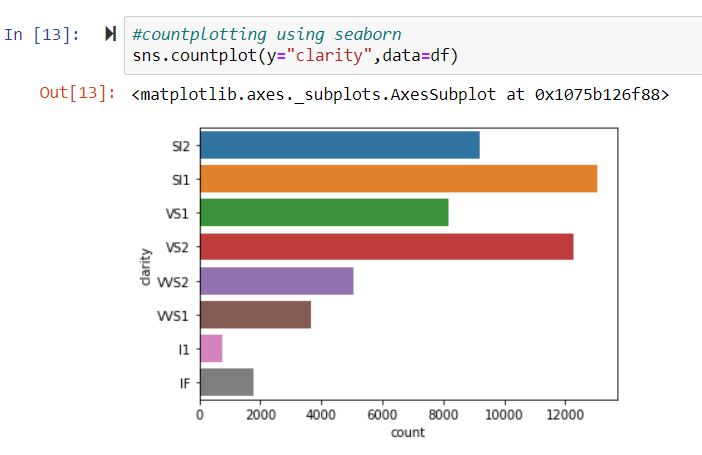
1. Scatter Plot

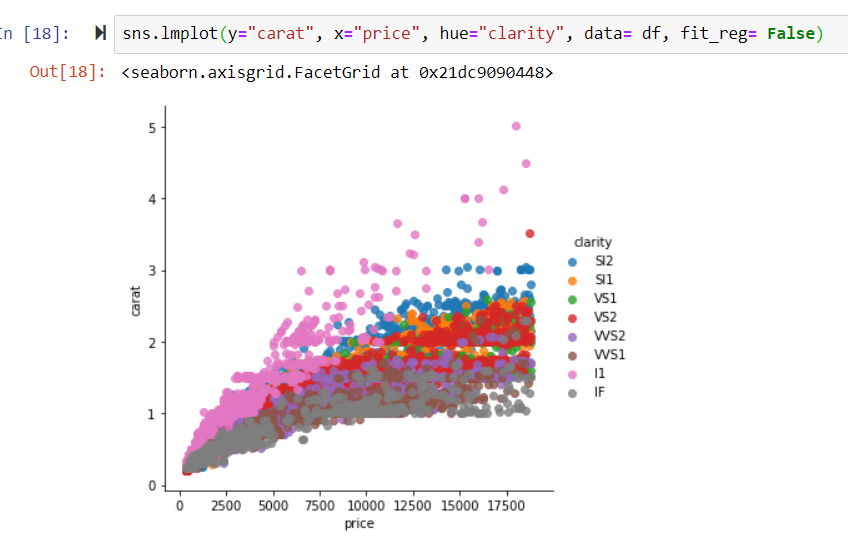


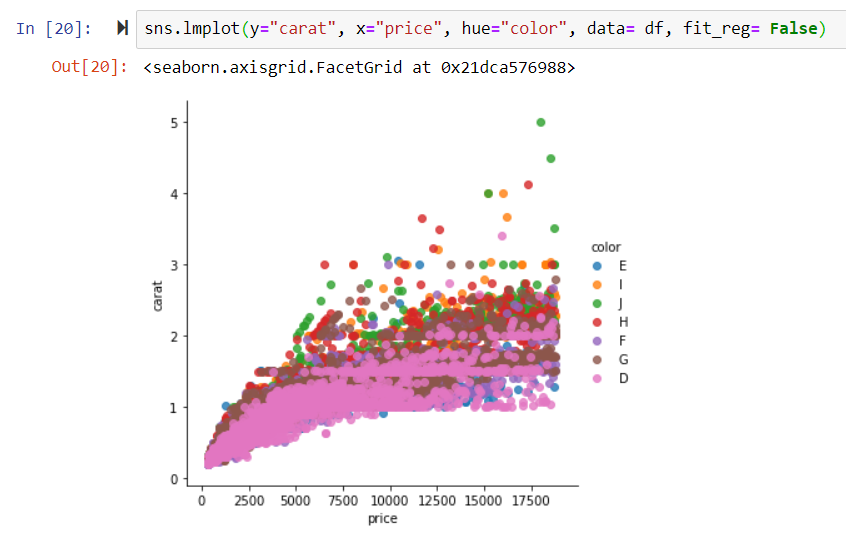
1. Axes3d



**Graphs we can use from seaborn library**

1. Countplot  
   
2. lmplot :-





We have plot the categorical data with price to see the variation using seaborn library

1. ***MODEL USED:-***

As now the data is fully ready to be trained by different algorithms of machine learning. To find the best accuracy we are making three models and find the best one by comparing the accuracy and mean squared error in both the three accuracy.

The three models that we are using to train the models are as follows:-

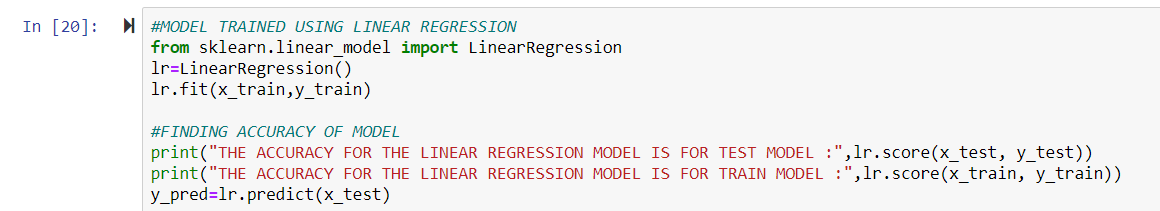
1. LINEAR REGRESSION:-

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other, but that there is some significant association between the two variables

A linear regression line has an equation of the form ***Y = a + b\*X***, where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

LEAST-SQUARE REGRESSION:-

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

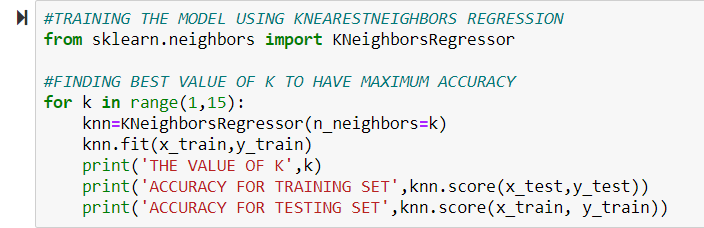


1. K NEAREST NEIGHBORS REGRESSOR:-

The *k*-nearest neighbors algorithm (*k*-NN) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis).[[1]](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm#cite_note-1) In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space). The output depends on whether *k*-NN is used for classification or regression:

* In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
* In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

The parameter here is most important is the value of k for which the value of accuracy is maximum i.e. k=5

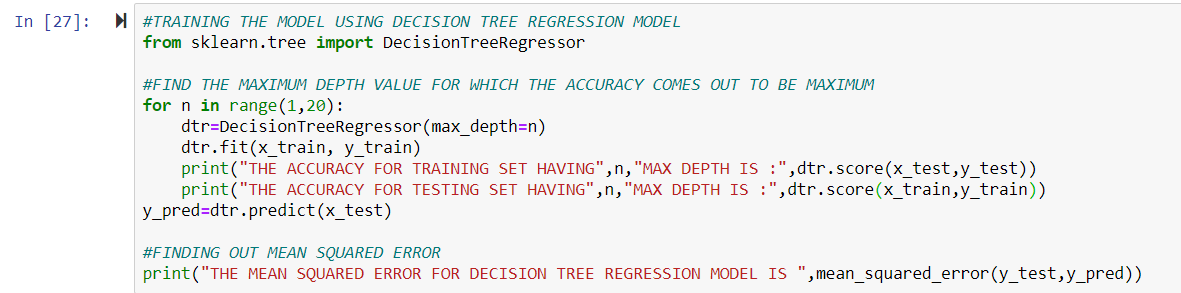


1. DECISION TREE REGRESSOR:-

A regression tree is similar to a classification tree, except that the Y variable takes ordered values and a regression model is fitted to each node to give the predicted values of Y. The AID and CART regression tree methods follow Algorithm 1, with the node impurity being the sum of squared deviations about the mean and the node predicting the sample mean of Y. This yields piecewise constant models. Although they are simple to interpret, the prediction accuracy of these models often lags behind that of models with more smoothness. It can be computationally impracticable, however, to extend this approach to piecewise linear models, because two linear models (one for each child node) must be fitted for every candidate split.

The parameter that is most important in this algorithm is max\_depth for which accuracy comes out to be maximum.

Here we have to find the best model from different regressor algorithms that will give the best accuracy as well as minimum mean squared error.



1. ***RESULTS OBTAINED (TILL DATE):-***

As we have trained the model with three different algorithm of machine learning the accuracy and mean squared error for three different algorithm are as follows :-

* 1. *LINEAR REGRESSION MODEL:-*

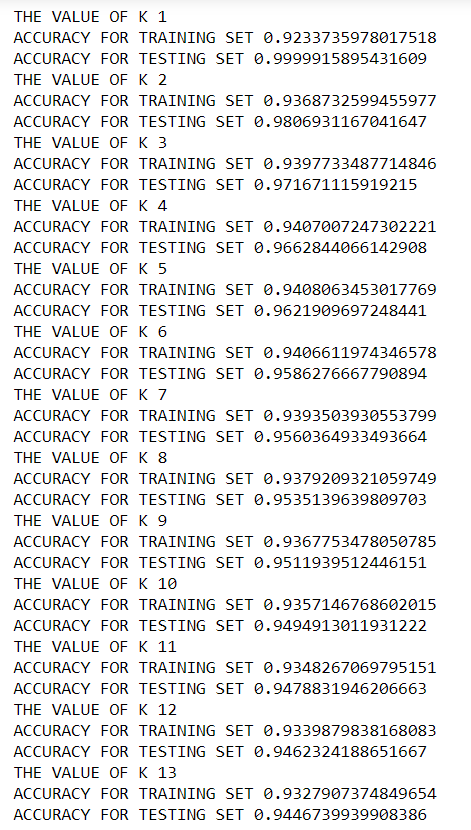
lra.PNG

ACCURACY FOR TESTING SET=92.12%

ACCURACY FOR TRAINING SET=91.94%

MEAN SQUARED ERROR =1248486.71

* 1. *K NEAREST NEIGHBORS REGRESSION MODEL:-*



BEST ACCURACY FOR TESTING SET=94.08%

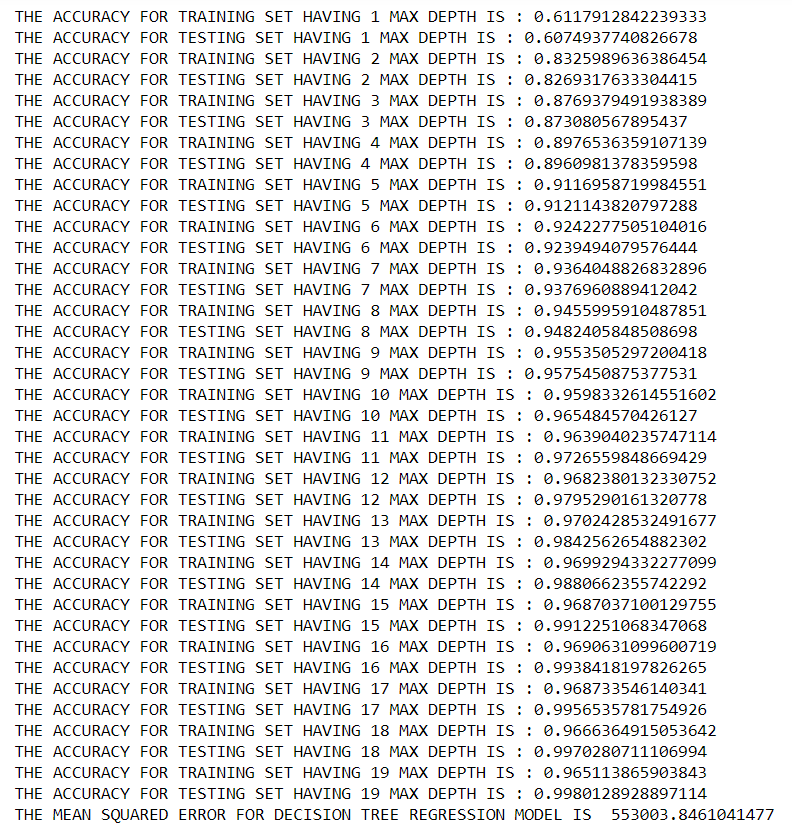
BEST ACCURACY FOR TRAINING SET=96.21%

MEAN SQUARED ERROR=1073849.65

BEST VALUE OF K=5

Here we have to find the best value of k which gives the best accuracy for the model.

* 1. *DECISION TREE REGRESSION MODEL:-*



ACCURACY FOR TESTING SET=97.024%

ACCURACY FOR TRAINING SET=98.42%

MEAN SQUARED ERROR=553003.84

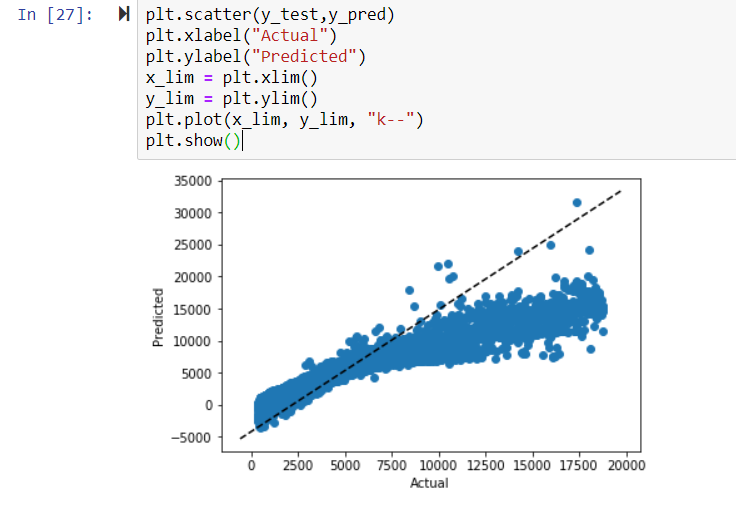
BEST VALUE OF MAXIMUM DEPTH=13

Here we have to find the best value of Maximum Depth which gives the best accuracy for the model.

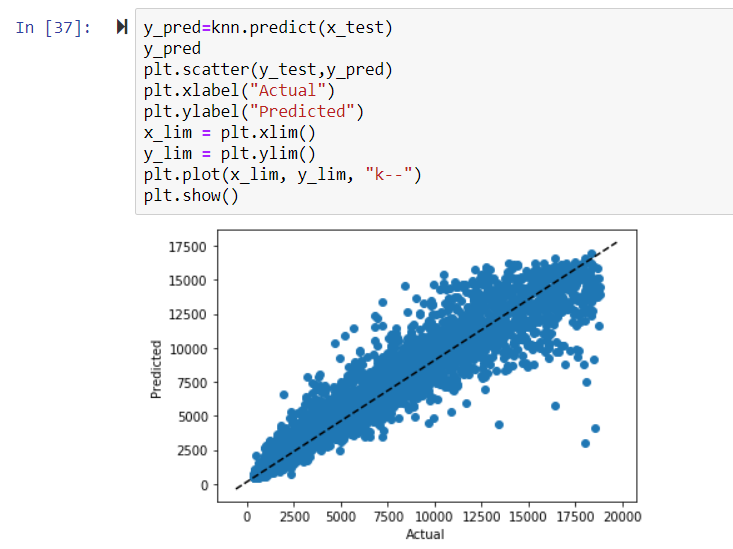
**So here the best model comes out to be model which is trained by the decision tree regression algorithm with parameter Maximum Depth taking value (13). The model gives out the max accuracy i.e. 97.024**

**The different fitting line for different model comes out to be:-**

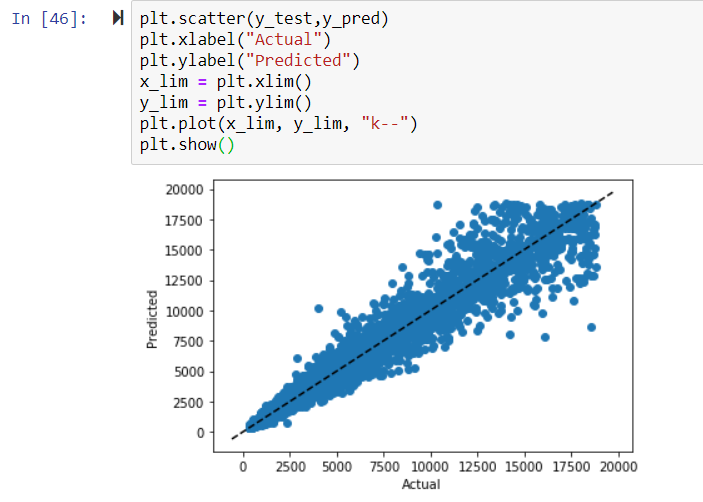
1. *LINEAR REGRESSION PLOT:-*



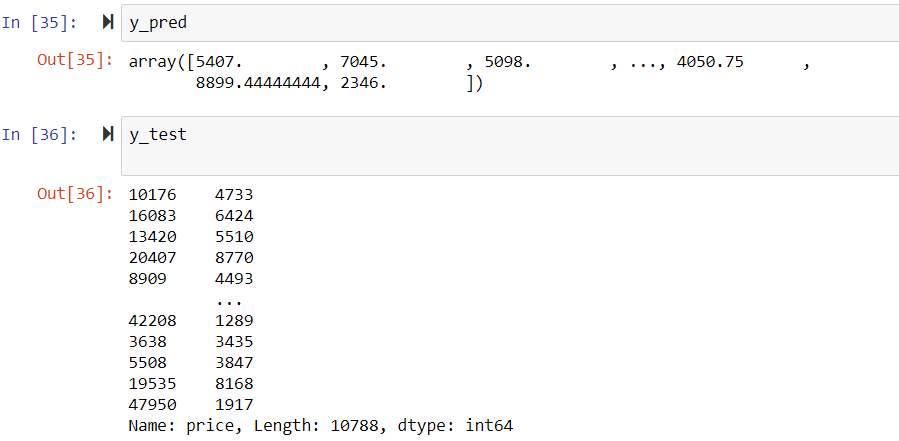
1. *K NEAREST NEIGHBORS REGRESSION PLOT:-*



1. *DECISION TREE REGRESSION:-*



And the value predicted by the model for the test dataset is:-



1. ***CONCLUSION:-***

So we investigated the effect of diamond’s special qualities like cut quality, weight, clarity etc on its price and developed a Decision Tree Regression model to decide the price dependence on these features.

The best model by comparing the accuracy and mean squared error is Decision Tree Regression Model. This Model predicts the price of diamonds dependency on their features. The maximum accuracy comes out to be 97.024% and minimum mean squared error comes out to be 553003.84.

Now we can predict the price of diamonds on this model by giving the diamond features to this model.