ABSTRACT

In this era of technology and innovation, human needs are rapidly increasing and new ways to improve the efficiency or to ease the process are constantly being invented. In this project we implement different machine learning models to predict the price of diamonds. This project is majorly based on a regression problem. Pre-processing of data is done in order to clean the data and miscellaneous data like diamonds with zero dimensions are removed.

The model predicts continuous values like the price of diamond.

We pick different details about the diamond from the dataset and apply various metrics to compare which model or algorithm works best, in order to use that in further studies.

PREFACE

1.1 INTRODUCTION:

- In this project various machine learning algorithms are implemented.
- This project is based on a regression problem.
- Here, we have used pre–processing on the data in order to clean the data.
- We have removed the miscellaneous data from the dataset like the diamonds which have zero dimensions.
- We have also removed the outliers in the data.
- It predicts continuous values like the price of diamonds.
- In this project, we get the different details about the diamond and predict the price of the diamond.
- We have used different metrics to compare which model was better than the rest to use that in our future endeavours.

1.2 Motivation of the project:

The project in hand works as a tool to predict whether or not we can predict the price of a diamond just by getting our hands on its features like colour, size, quality etc. This was also aimed at checking the accuracy of the algorithm and of the project as a whole.

The greater success and accuracy of the project will help us to implement it in our real-life, day-to-day practices.

1.3 Basic description of the project:

In this project, we implement different machine-learning models to predict the price of diamonds. This project is majorly based on a regression problem. Pre-processing of data is done in order to clean the data and miscellaneous data like diamonds with zero dimensions are removed. The model predicts continuous values like the price of a diamond.

We pick different details about the diamond from the dataset and apply various metrics to compare which model or algorithm works best, in order to use that in further studies.

Literature review

2.1 General

This project uses different machine learning algorithms and saves the better models and then compares the models to see which model gives a more accurate prediction about the price of the diamond.

2.2 Review of related works

There are other places where this type of work has been done but they rely on only one algorithm to predict the price like only XGboost that gave good price predictions.

Related Theories and Algorithms.

Fundamental theories underlying the work:

1. What is RMSE?

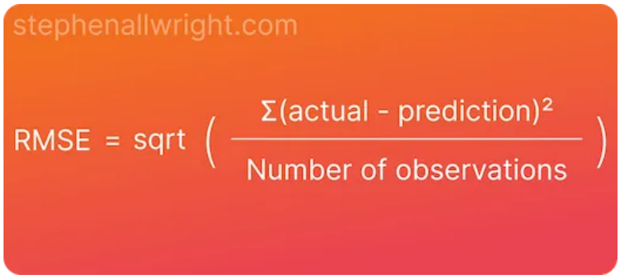
Root Mean Squared Error (RMSE) is the square root of the mean squared error between the predicted and actual values.

Squared error, also known as L2 loss, is a row-level error calculation where the difference between the prediction and the actual is squared. RMSE is the aggregated mean and subsequent square root of these errors, which helps us understand the model performance over the whole dataset.

A benefit of using RMSE is that the metric it produces is on the same scale as the unit being predicted. For example, calculating RMSE for a house price prediction model would give the error in terms of house price, which can help end users easily understand model performance.

RMSE mathematical formula

The formula for calculating RMSE is:



stephenallwright.com

$$\text{RMSE} = \text{sqrt}\left(\frac{\Sigma(\text{actual} - \text{prediction})^2}{\text{Number of observations}}\right)$$

RMSE value interpretation

The closer RMSE is to 0, the more accurate the model is. But RMSE is returned on the same scale as the target you are predicting and therefore there isn't a general rule for how to interpret ranges of values. The interpretation of your value can only be evaluated within your dataset.

2. What is MSE?

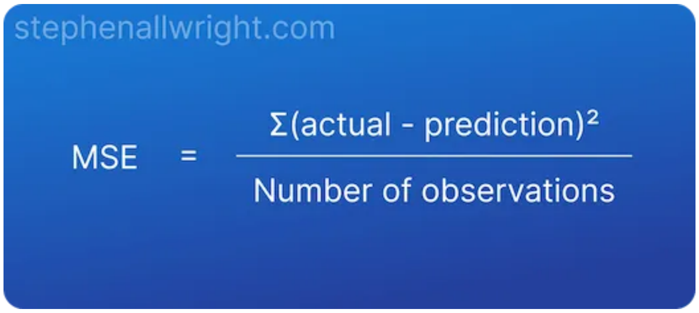
Mean Squared Error (MSE) is the average squared error between actual and predicted values.

Squared error, also known as L2 loss, is a row-level error calculation where the difference between the prediction and the actual is squared. MSE is the aggregated mean of these errors, which helps us understand the model performance over the whole dataset.

The main draw for using MSE is that it squares the error, which results in large errors being punished or clearly highlighted. It's therefore useful when working on models where occasional large errors must be minimised.

MSE mathematical formula

The formula for calculating MSE is:



$$\text{MSE} = \frac{\Sigma(\text{actual - prediction})^2}{\text{Number of observations}}$$

How to interpret MSE?

MSE should be interpreted as an error metric where the closer your value is to 0, the more accurate your model is. However, MSE is simply the average of the squared errors, meaning the resulting value will unfortunately not be understood within the context of your model target.

3. What is MAE?

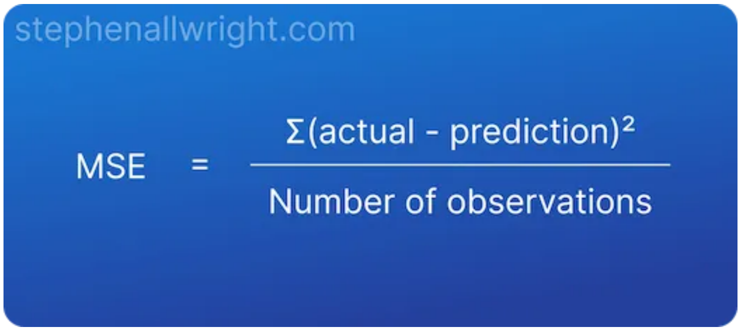
MAE (Mean Absolute Error) is the average absolute error between actual and predicted values.

Absolute error, also known as L1 loss, is a row-level error calculation where the non-negative difference between the prediction and the actual is calculated. MAE is the aggregated mean of these errors, which helps us understand the model performance over the whole dataset.

MAE is a popular metric to use as the error value is easily interpreted. This is because the value is on the same scale as the target you are predicting for.

MAE mathematical formula

The formula for calculating MAE is:



stephenallwright.com

$$\text{MSE} = \frac{\Sigma(\text{actual - prediction})^2}{\text{Number of observations}}$$

How to interpret MAE?

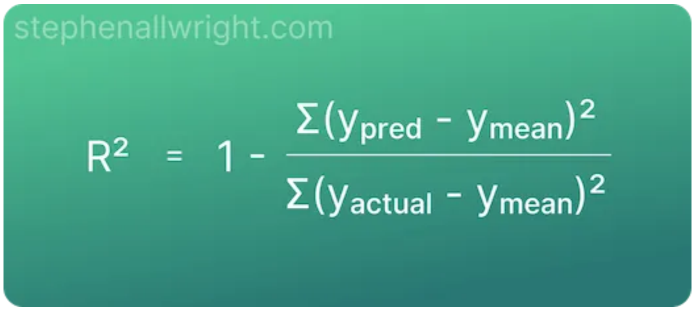
The closer MAE is to 0, the more accurate the model is. But MAE is returned on the same scale as the target you are predicting for and therefore there isn’t a general rule for how to interpret ranges of values. The interpretation of your value can only be evaluated within your dataset.

4. What is R Squared?

R Squared (also known as R2) is a metric for assessing the performance of regression machine learning models. Unlike other metrics, such as MAE or RMSE, it is not a measure of how accurate the predictions are, but instead a measure of fit. **R Squared measures how much of the dependent variable variation is explained by the independent variables in the model.**

R Squared mathematical formula

The formula for calculating R Squared is as follows:


stephenallwright.com

$$R^2 = 1 - \frac{\Sigma(y_{pred} - y_{mean})^2}{\Sigma(y_{actual} - y_{mean})^2}$$

How to interpret R Squared?

R Squared can be interpreted as the percentage of the dependent variable variance which is explained by the independent variables. Put simply, it measures the extent to which the model features can be used to explain the model target.

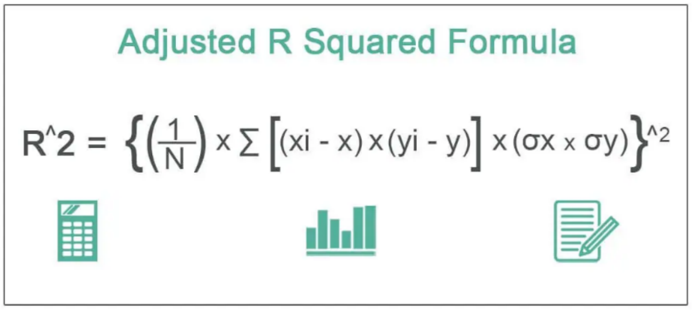
For example, an R Squared value of 0.9 would imply that 90% of the target variance can be explained by the model features, whilst a value of 0.2 would suggest that the model features are only able to account for 20% of the variance.

What is Adjusted R Squared?

Adjusted R Squared or Modified R^2 determines the extent of the variance of the dependent variable, which the independent variable can explain. The speciality of the modified R^2 is that it does not consider the impact of all independent variables but only those which impact the variation of the dependent variable. Therefore, the value of the modified R^2 can also be negative, though it is not always negative.

R Squared mathematical formula

The formula for calculating R adjusted Squared is as follows:



Adjusted R Squared Formula

$$R\hat{}2 = \left\{\left(\frac{1}{N}\right) x \sum \left[(xi - x) x (yi - y)\right] x (\sigma x \times \sigma y)\right\}^{\hat{}2}$$

- R^2= adjusted R square of the **regression equation**
- N= Number of observations in the regression equation
- Xi= Independent variable of the regression equation
- X= Mean of the independent variable of the regression equation
- Yi= Dependent variable of the regression equation
- Y= **Mean** of the dependent variable of the regression equation
- σx = Standard deviation of the independent variable
- σy = Standard deviation of the dependent variable.

How to interpret adjusted R Squared?

Adjusted R Square is a significant output to determine whether the data set is a good fit. Someone does a regression equation to validate whether what he thinks of the relationship between two variables is also validated by the regression equation. The higher the value, the better the regression equation, which implies that the independent variable chosen to determine the dependent variable is chosen appropriately. Ideally, a researcher will look for the coefficient of determination closest to 100%.

Related Theories and Algorithms

Fundamental algorithms:

1. Linear regression

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.

2. Decision Tree

Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

3. Random Forest

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and predicts the final output**.**

4. K-NN algorithm

The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.

The k-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.

5. XGBoost

XGBoost is an implementation of Gradient Boosted decision trees. XGBoost models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers /predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

The Data Visualizations used for the analysis.

Different kinds of plots used for data interpretation.

1. Kernel Density Estimation (KDE) Plot:

KDE plots estimate the probability density function of continuous data. They create a smooth curve by placing kernels at data points. KDE plots help visualize data distribution, compare multiple distributions, and identify patterns or anomalies. They are useful in exploratory data analysis and can be implemented in Python or R using libraries like seaborn or ggplot2.

2. Pairplot:

A pairplot is a visualization tool used to explore the relationships between multiple variables. It displays scatterplots for pairwise combinations of variables in a dataset and histograms for individual variables. It provides a quick overview of the relationships and distributions in the data.

3. Heatmap:

A heatmap is a visual representation of data that uses colors to indicate the values of a matrix or table. It is commonly used to display the relationship or intensity of variables in a dataset. Each cell in the heatmap is filled with a color corresponding to the value it represents, making it easy to identify patterns and trends. Heatmaps are often used in various fields such as data analysis, biology, and finance. They provide a concise and intuitive way to visualize complex data and aid in decision-making processes.

4.2D Kernel Density Estimation (KDE) plot:

In general, a 2D Kernel Density Estimation (KDE) plot is a visualization technique that represents the estimated probability density of a two-dimensional dataset. It is used to explore the relationship between two continuous variables. The plot is created by placing a kernel (a mathematical function) at each data point and summing them up to create a smooth, continuous surface. The density of the data is represented using a colormap, where higher densities are typically shown in darker shades and lower densities in lighter shades.

A 2D KDE plot provides insights into the distribution and concentration of data points in a two-dimensional space. It can reveal patterns, clusters, and areas of high or low density in the data, helping to identify relationships or trends between the variables. It is often used in exploratory data analysis and can be created using various programming libraries, such as seaborn in Python.

5. Scatter plot

A scatter plot is a visual representation of the relationship between two variables in a dataset. It displays individual data points as dots on a two-dimensional graph, with one variable plotted on the x-axis and the other on the y-axis. Each dot represents a data point with a specific value for both variables.The scatter plot helps to identify the nature of the relationship between the variables. It can show whether the variables are positively or negatively correlated, or if there is no significant relationship between them. Patterns such as clusters, trends, or outliers can be observed in the scatter plot.

By examining the scatter plot, one can make assessments about the strength and direction of the relationship, identify potential patterns or trends, and detect any potential outliers or unusual data points. Scatter plots are widely used in data analysis and are helpful for understanding the relationships between variables and making informed decisions based on the observed patterns.

6. Regression line plot:

A regression line plot is a visual representation of the linear relationship between two variables. It consists of a scatter plot of the data points and a line that best fits the data. The line represents the trend and direction of the relationship. It helps determine if the relationship is positive or negative and the rate of change between the variables. The plot allows for predictions and estimation of the dependent variable based on the independent variable. It aids in understanding the relationship between variables and making informed decisions or predictions.

7. Box plot:

A box plot is a visual summary of the distribution of a dataset. It consists of a box that represents the interquartile range, with the median indicated inside. The whiskers extend to show the range of the data, excluding outliers. Box plots help compare distributions, identify skewness, and detect outliers. They provide insights into the central tendency and variability of the data. By examining a box plot, one can quickly understand the data's spread, central values, and potential outliers. They are widely used in exploratory data analysis to make comparisons between groups or variables, especially with large datasets.

8. Violin plot:

A violin plot is a visual representation of the distribution and density of a dataset. It combines a box plot with a kernel density plot, providing a comprehensive summary of the data's characteristics.In a violin plot, each violin represents a variable or group, and its width corresponds to the density of data at different values. The thicker sections indicate regions of higher data density. Inside the violin, a white dot represents the median, and optionally, a box plot can be included.

Violin plots are effective for comparing distributions across different categories or variables. They provide insights into the shape, spread, and skewness of the data. They can also reveal multimodal distributions or asymmetrical patterns. Violin plots are commonly used in exploratory data analysis to understand the distributional characteristics of a dataset and make comparisons between groups or variables.

9. Joint plot:

A jointplot is a type of plot that combines multiple visualizations to explore the relationship between two variables. It is particularly useful for understanding the joint distribution, correlation, and marginal distributions of the variables.The jointplot typically includes a scatter plot, which shows the individual data points of the two variables. Additionally, it may include a regression line to visualize the linear relationship between the variables. Along the margins of the plot, the distributions of each variable are displayed using histograms or kernel density plots.

This plot provides a comprehensive view of the relationship between the variables, allowing for the assessment of correlation, linearity, and the concentration of data points. It helps identify patterns, outliers, and any nonlinear relationships between the variables.

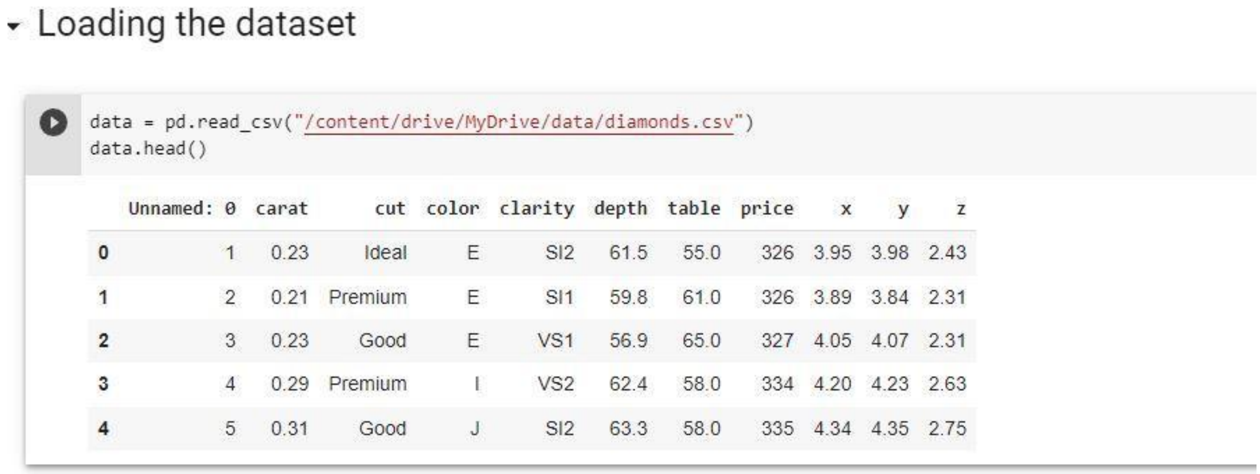
The jointplot is widely used in exploratory data analysis to gain insights into the relationship between two variables and to assess their distributions. It helps in making informed decisions and understanding the nature of the data.

Simulation Results

Experimental setup:

In this project, we have used a diamond dataset that was taken from Kaggle website, we have used different Python libraries that were important for loading different machine learning models then, we have loaded the dataset into a data frame variable, then we tried to delete useless data and remove the outliers and then we visualise the data using graphs and heat maps. We split the whole data into train data which is used for training the model and test data which is used for testing the model. And created pipelines for using different machine learning algorithms like linear regression, k-nearest neighbour, decision tree random forest and XGBoost.

Dataset Used in the project preview

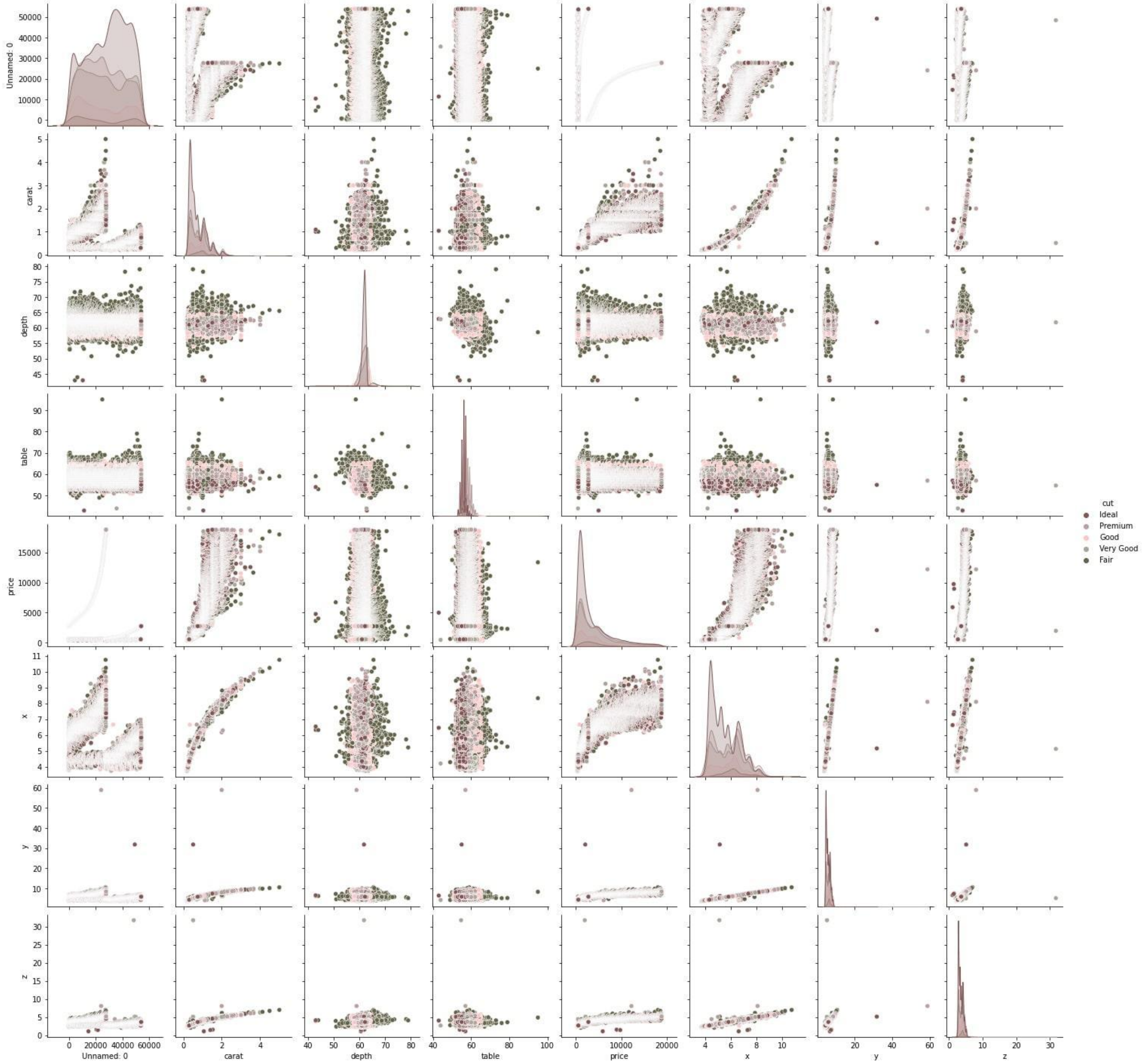


Loading the dataset

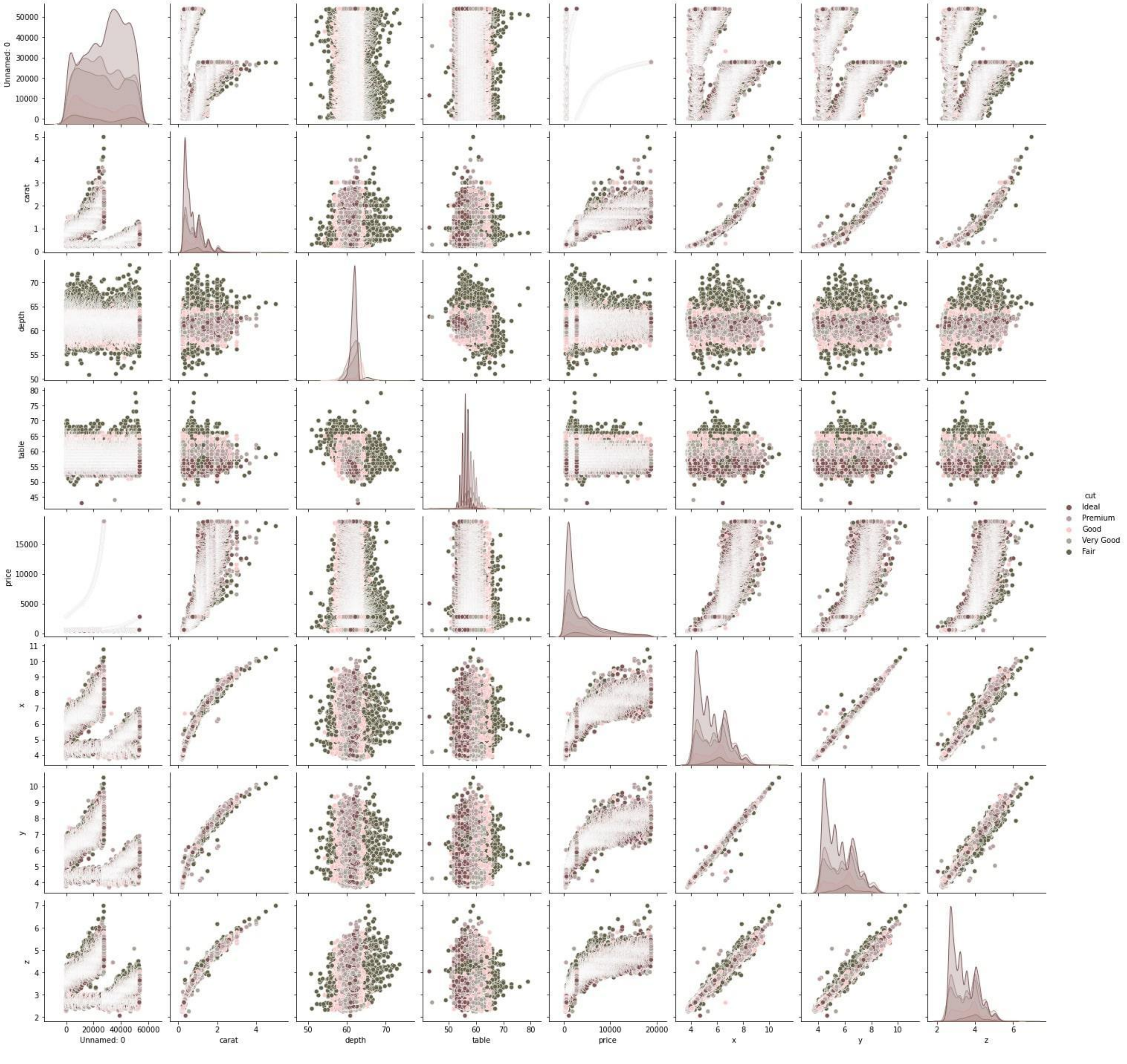
```
data = pd.read_csv("/content/drive/MyDrive/data/diamonds.csv")
data.head()
```

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

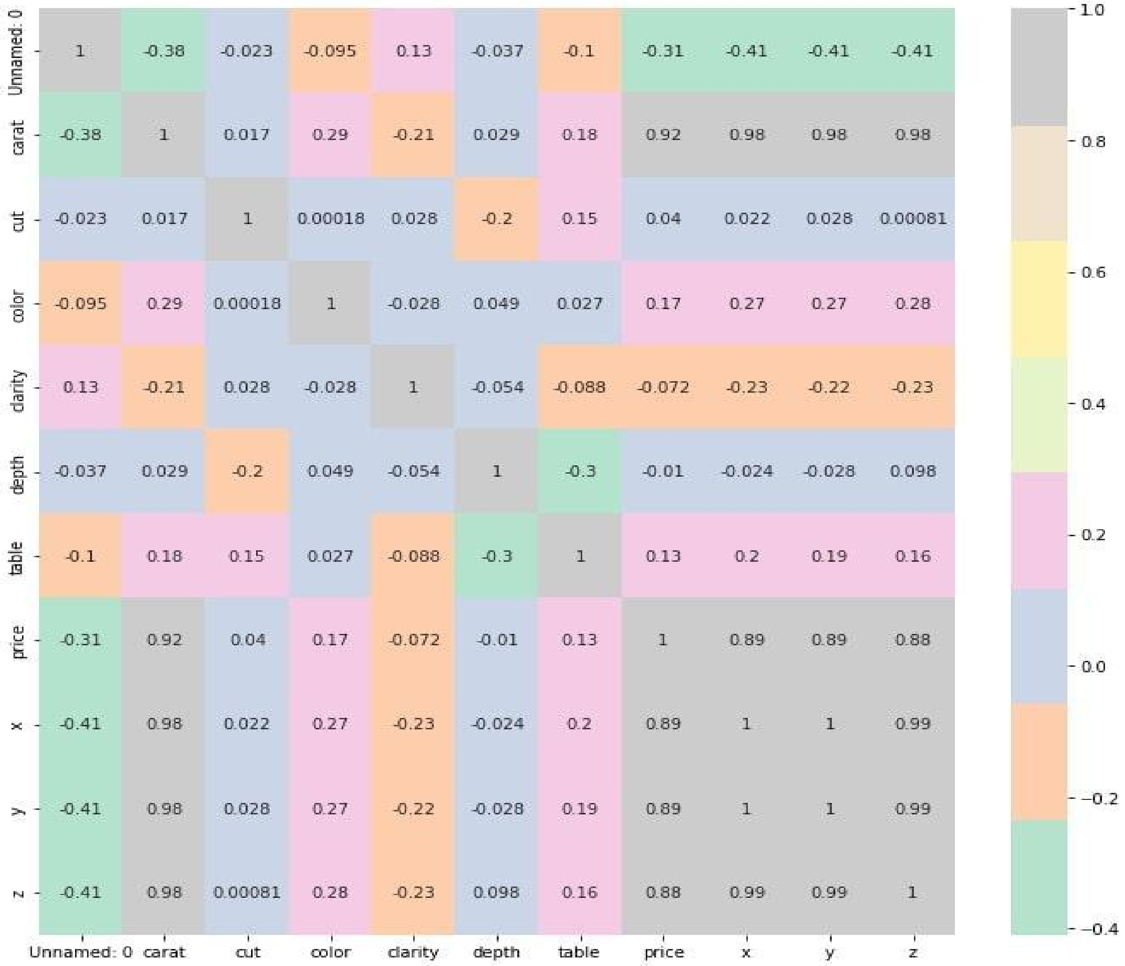
The Pairplot of the dataset before cleaning the outliers.



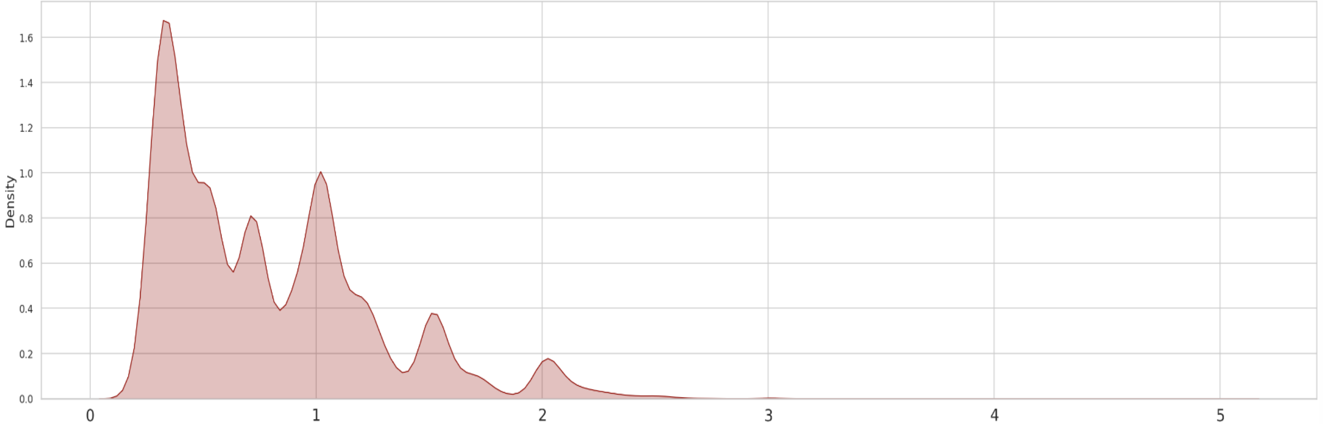
The Pairplot of the dataset after cleaning the outliers.



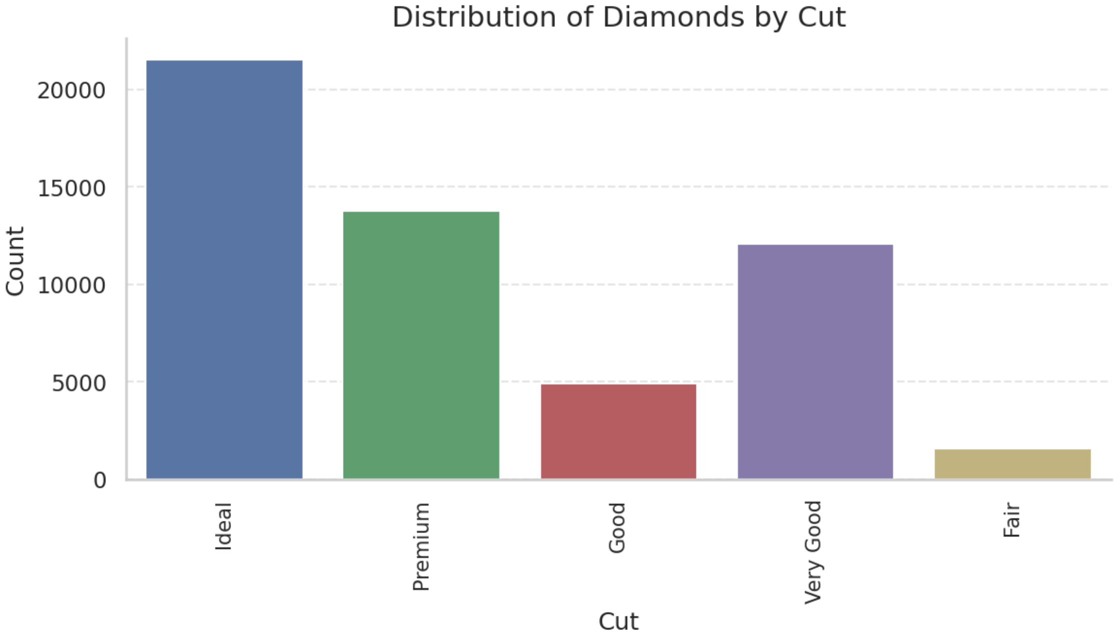
Heatmap of the different features of the dataset.



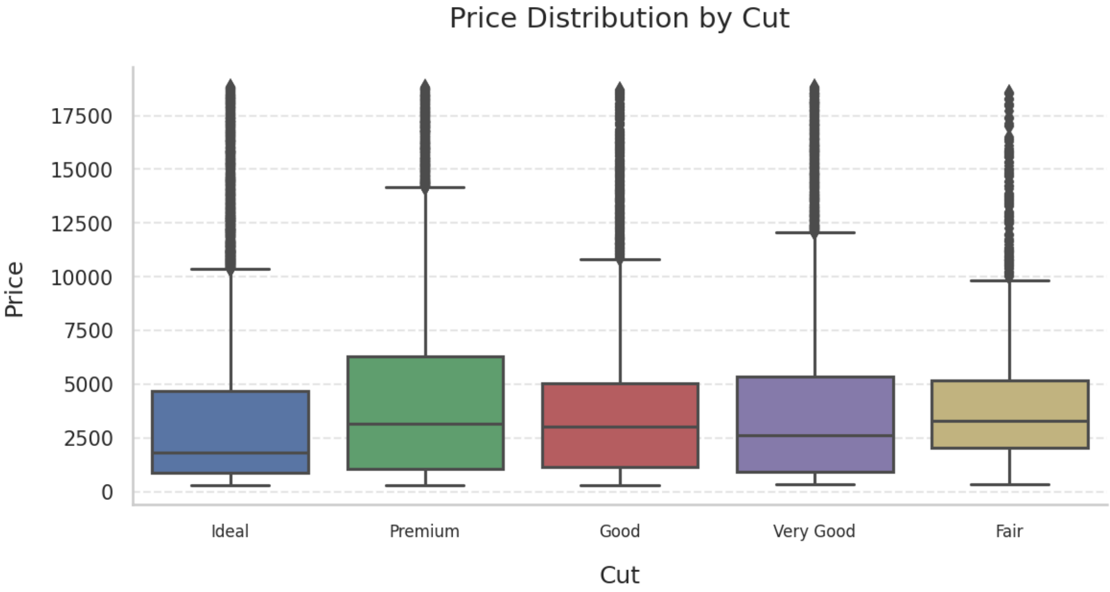
Carat KDE plot.



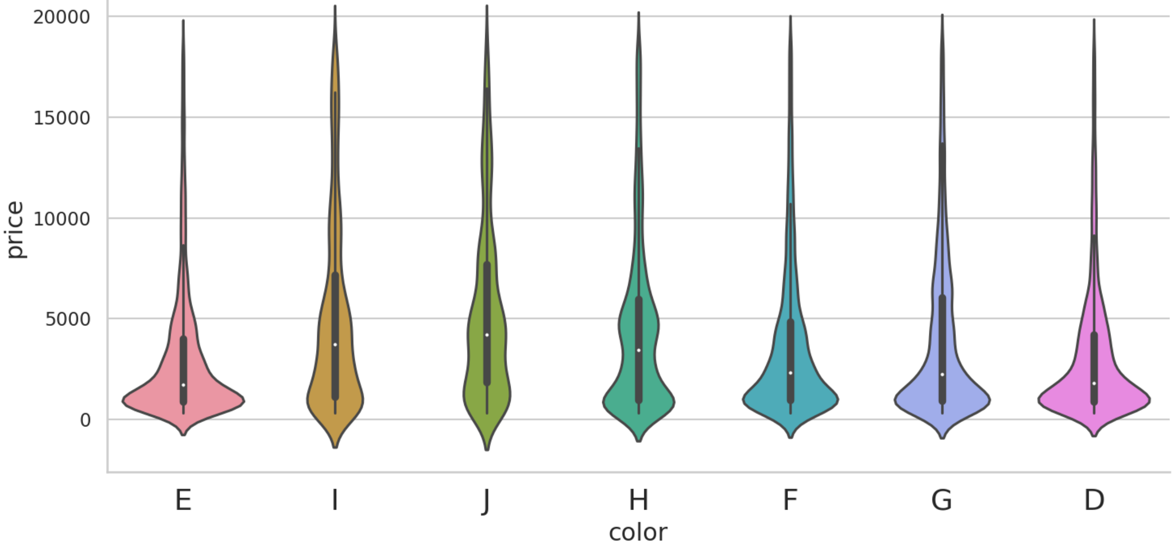
Cut barplot.



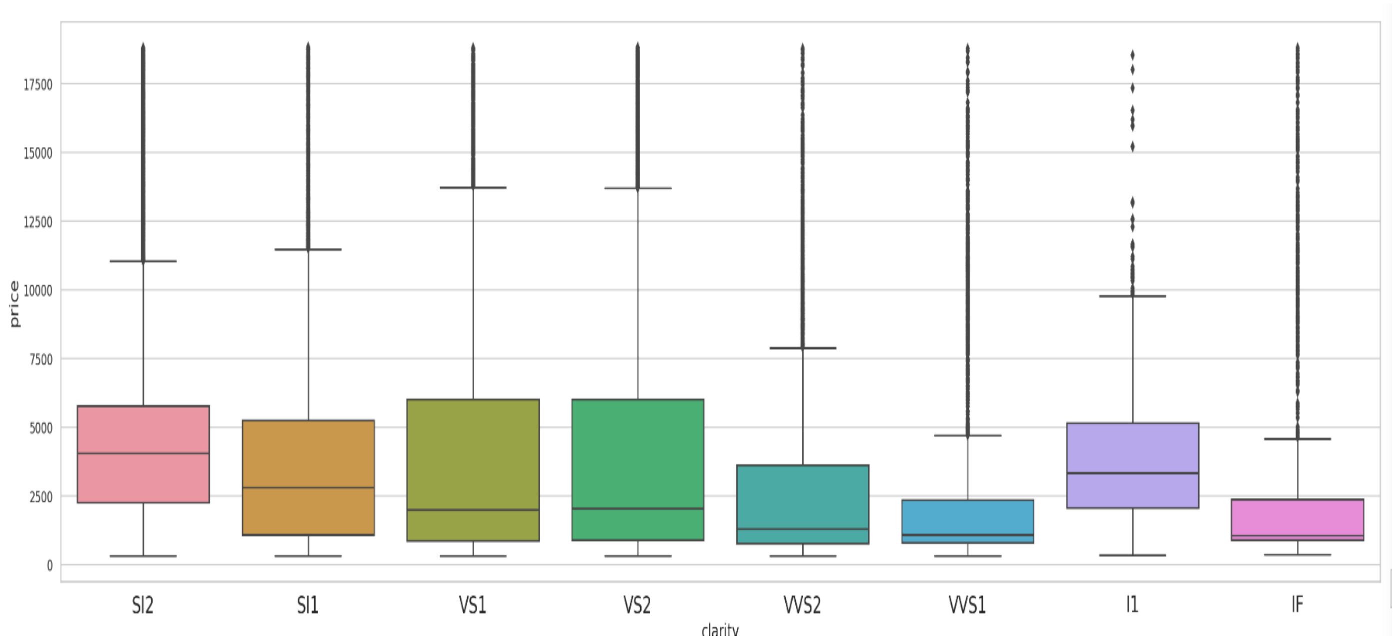
Cut Boxplot.



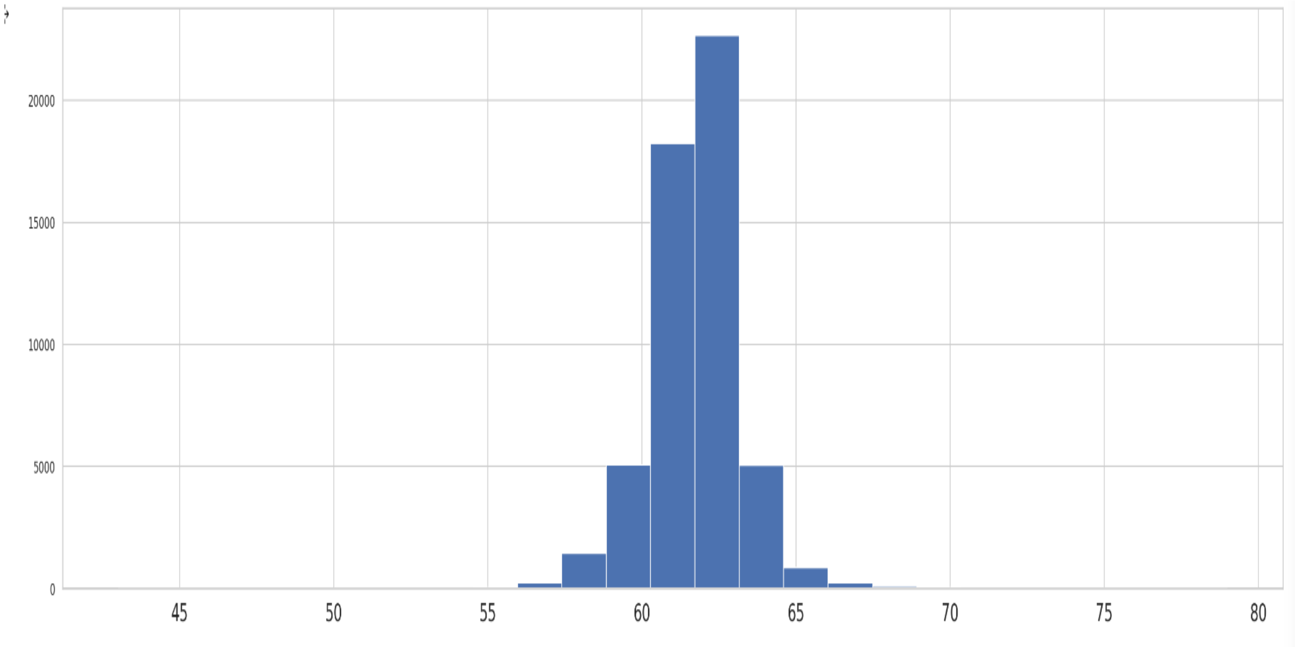
Color ViolinPlot.



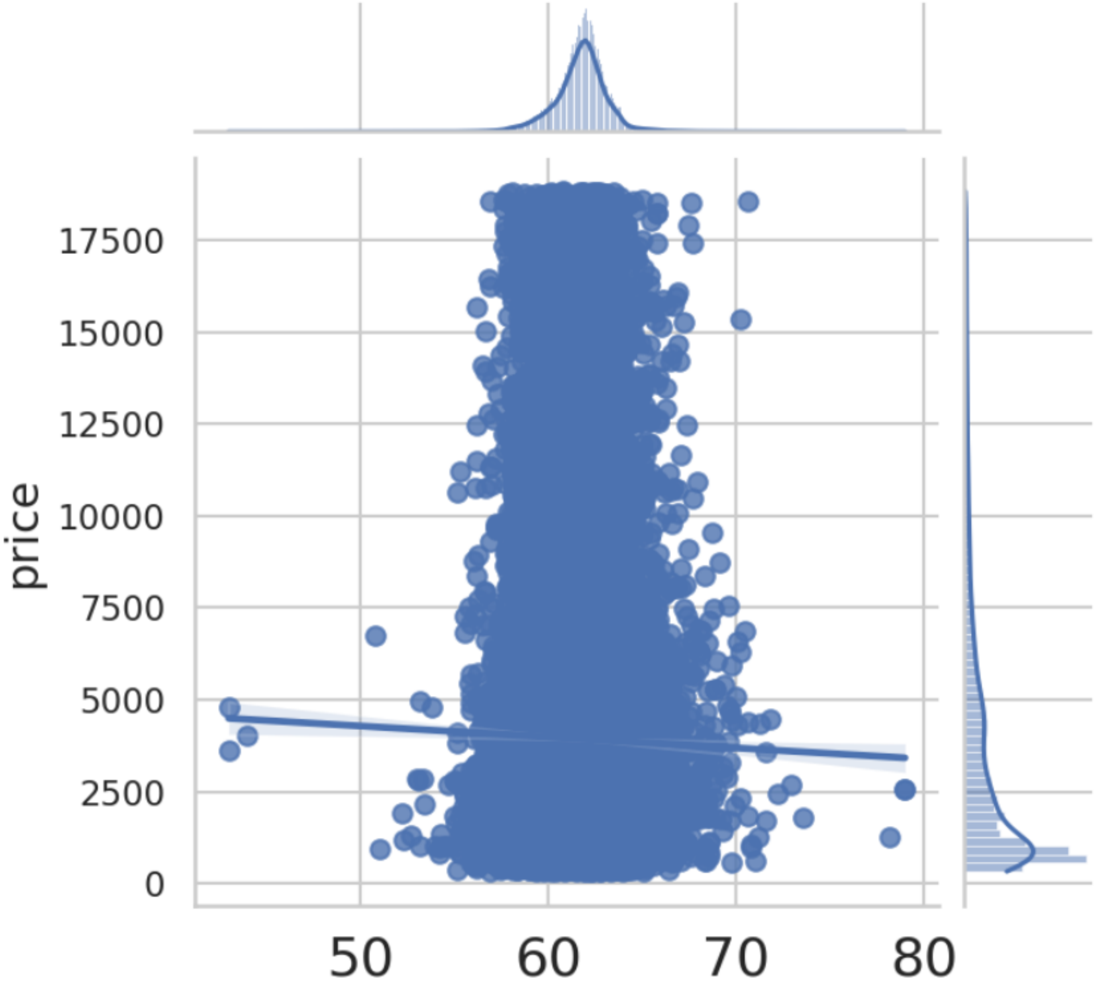
Clarity BoxPlot.



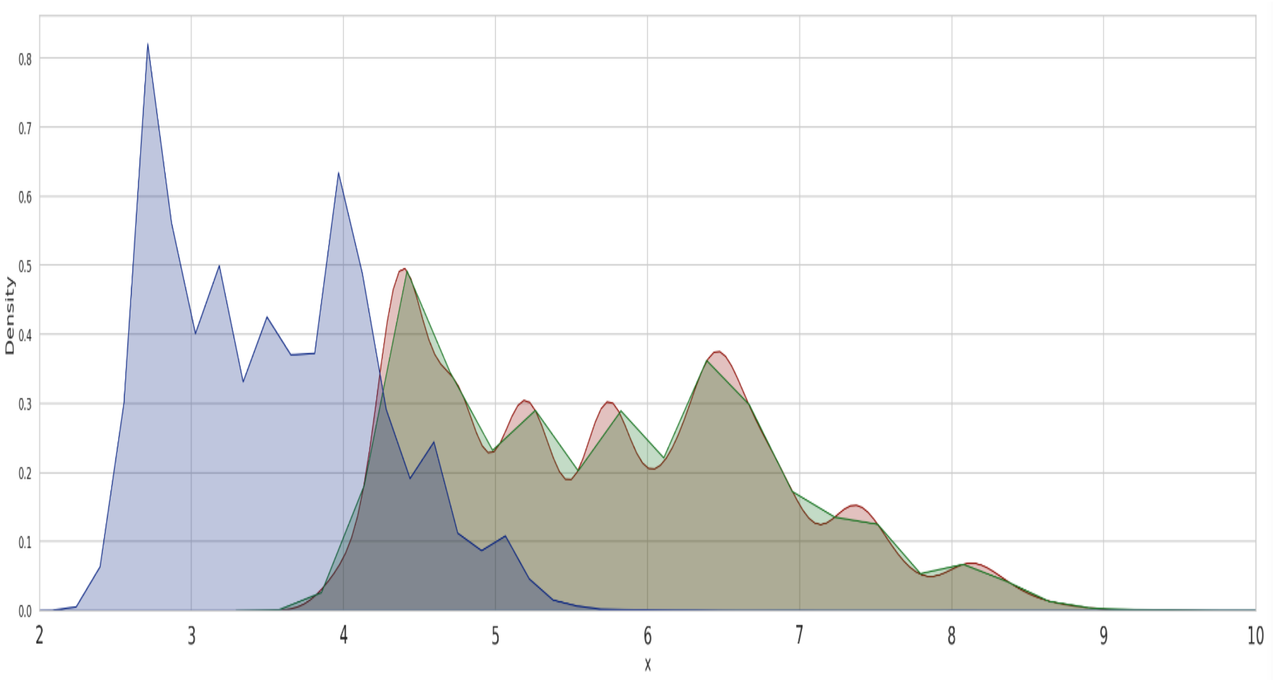
Depth Histogram.



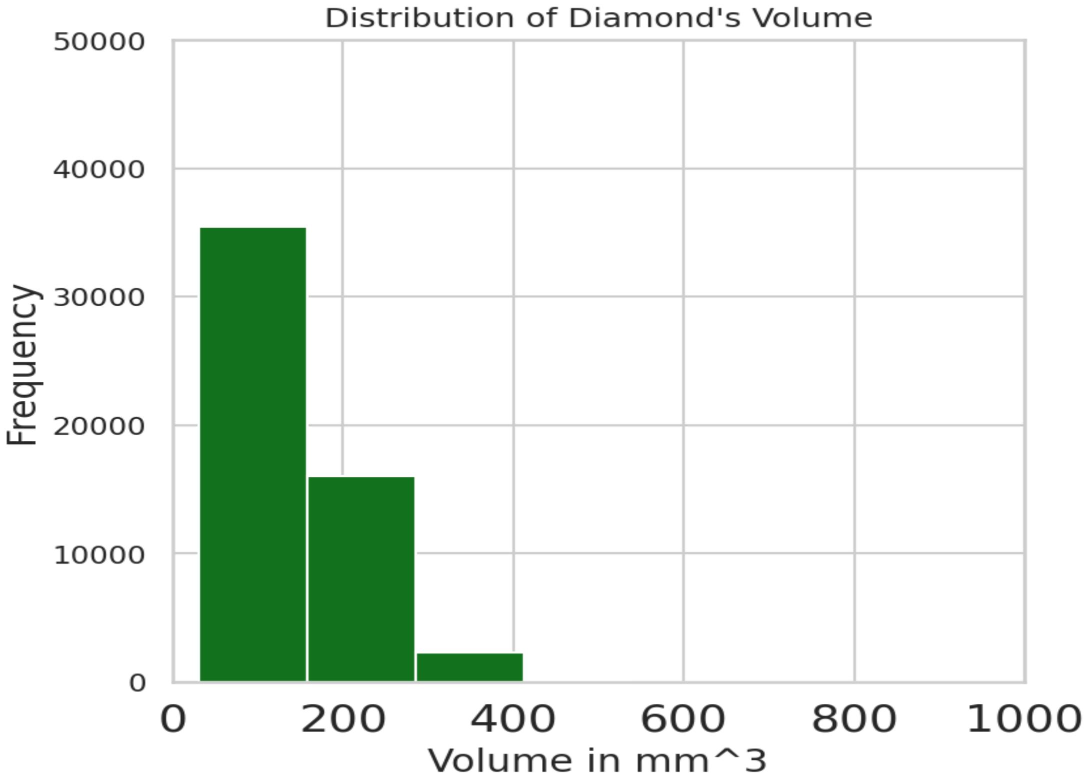
Depth and Price JointPlot



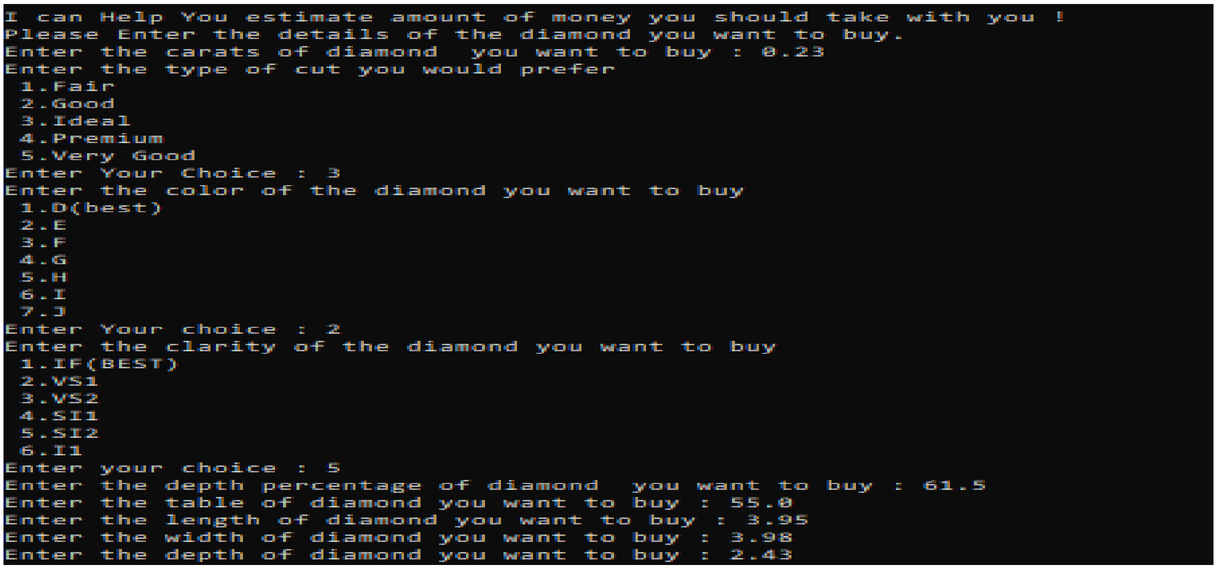
Dimensions KDE plot.



Volume Distribution plot.



The application Program of the Diamond prediction

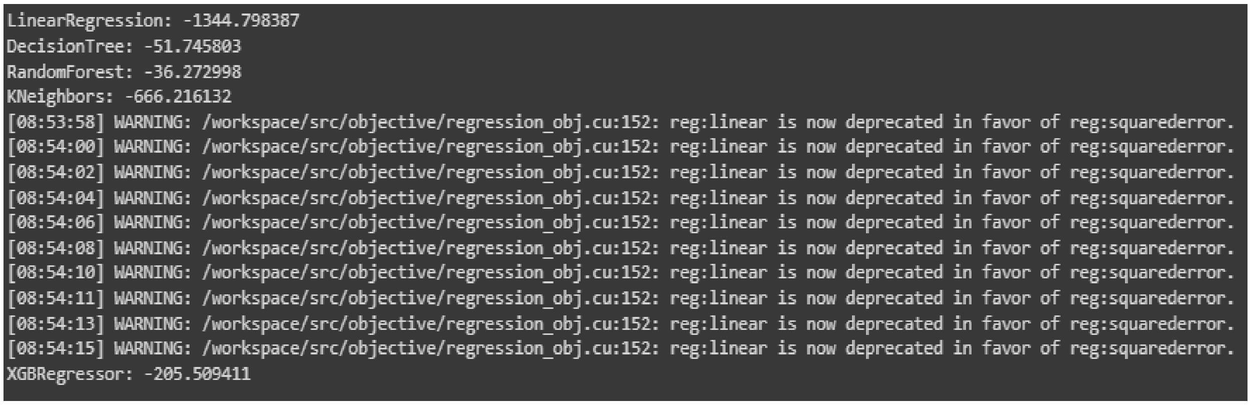


```
I can Help You estimate amount of money you should take with you !
Please Enter the details of the diamond you want to buy.
Enter the carats of diamond  you want to buy : 0.23
Enter the type of cut you would prefer
 1.Fair
 2.Good
 3.Ideal
 4.Premium
 5.Very Good
Enter Your Choice : 3
Enter the color of the diamond you want to buy
 1.D(best)
 2.E
 3.F
 4.G
 5.H
 6.I
 7.J
Enter Your choice : 2
Enter the clarity of the diamond you want to buy
 1.IF(BEST)
 2.VS1
 3.VS2
 4.SI1
 5.SI2
 6.I1
Enter your choice : 5
Enter the depth percentage of diamond  you want to buy : 61.5
Enter the table of diamond you want to buy : 55.0
Enter the length of diamond you want to buy : 3.95
Enter the width of diamond you want to buy : 3.98
Enter the depth of diamond you want to buy : 2.43
```

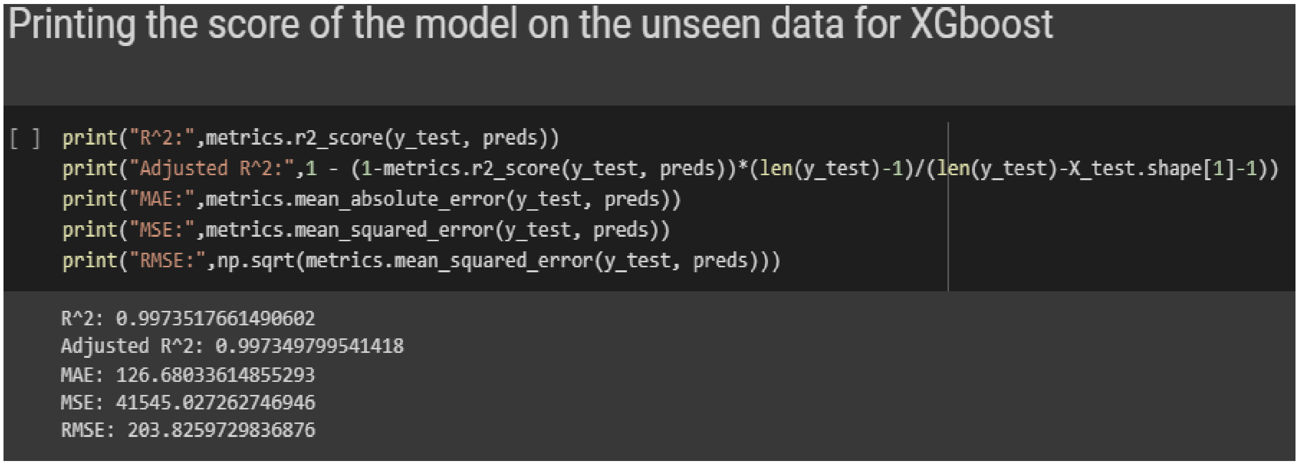
Simulation Results

Experimental results:

After training different models of machine learning algorithms like linear regression, decision tree, random forest, k nearest neighbor, and XGBoost we saw that random forest and decision tree were giving the best results even though XGBoost was a more advanced model but it's not working for this model. But after further analysis, we found that even though the scores of the random forest were very good but decision tree predicts more accurate results than the random forest. Hence we use the decision tree model for predicting the price of diamonds.We have made an application which uses this model to predict the price of diamond using different features entered by the user.



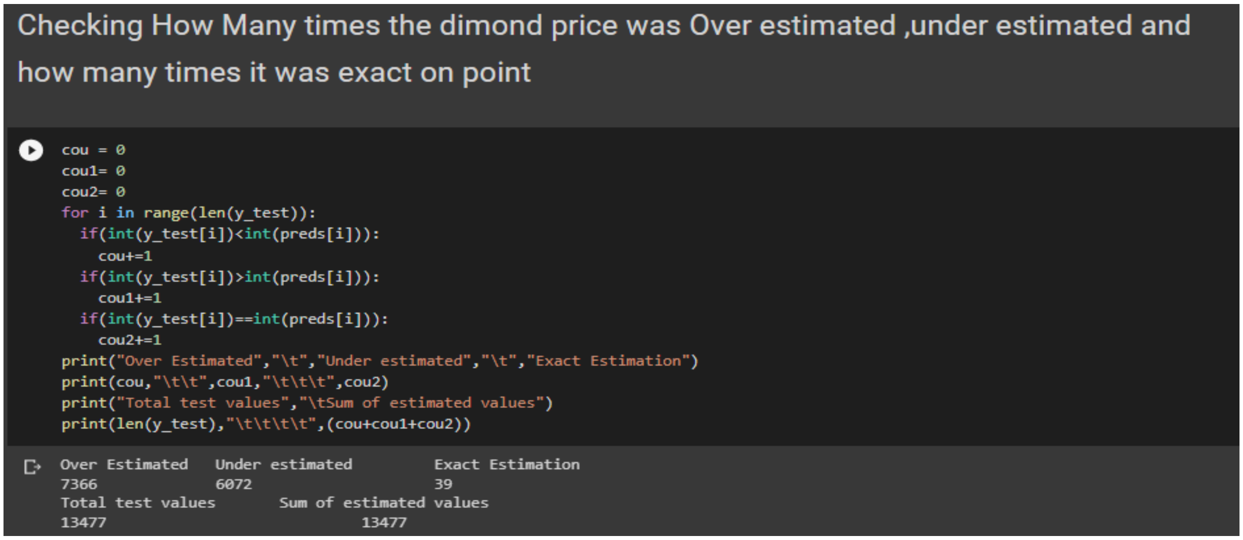
```
LinearRegression: -1344.798387
DecisionTree: -51.745803
RandomForest: -36.272998
KNeighbors: -666.216132
[08:53:58] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:00] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:02] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:04] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:06] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:08] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:10] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:11] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:13] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
[08:54:15] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
XGBRegressor: -205.509411
```



Printing the score of the model on the unseen data for XGboost

```
[ ]  print("R^2:",metrics.r2_score(y_test, preds))
     print("Adjusted R^2:",1 - (1-metrics.r2_score(y_test, preds))*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1))
     print("MAE:",metrics.mean_absolute_error(y_test, preds))
     print("MSE:",metrics.mean_squared_error(y_test, preds))
     print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, preds)))
```

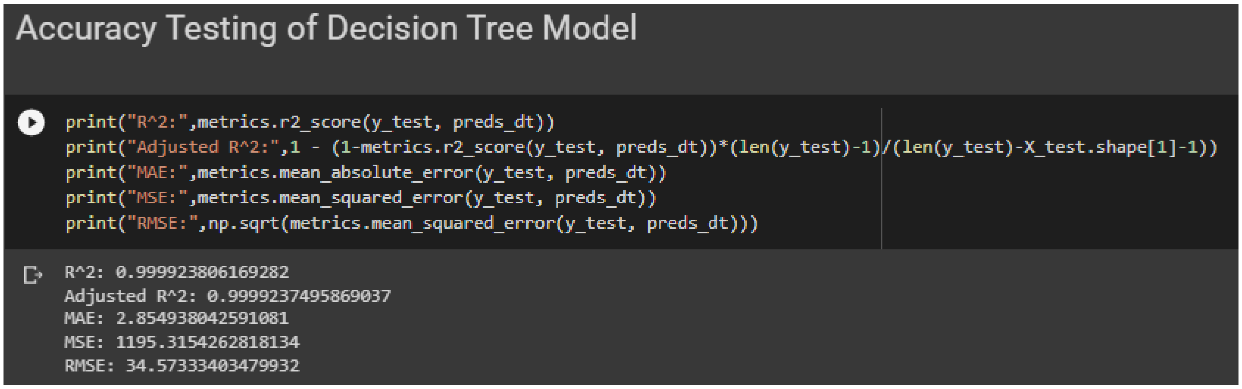
```
     R^2: 0.9973517661490602
     Adjusted R^2: 0.997349799541418
     MAE: 126.68033614855293
     MSE: 41545.027262746946
     RMSE: 203.8259729836876
```



Checking How Many times the dimond price was Over estimated ,under estimated and how many times it was exact on point

```
cou = 0
cou1= 0
cou2= 0
for i in range(len(y_test)):
  if(int(y_test[i])<int(preds[i])):
    cou+=1
  if(int(y_test[i])>int(preds[i])):
    cou1+=1
  if(int(y_test[i])==int(preds[i])):
    cou2+=1
print("Over Estimated","\t","Under estimated","\t","Exact Estimation")
print(cou,"\t\t",cou1,"\t\t\t",cou2)
print("Total test values","\tSum of estimated values")
print(len(y_test),"\t\t\t\t",(cou+cou1+cou2))
```

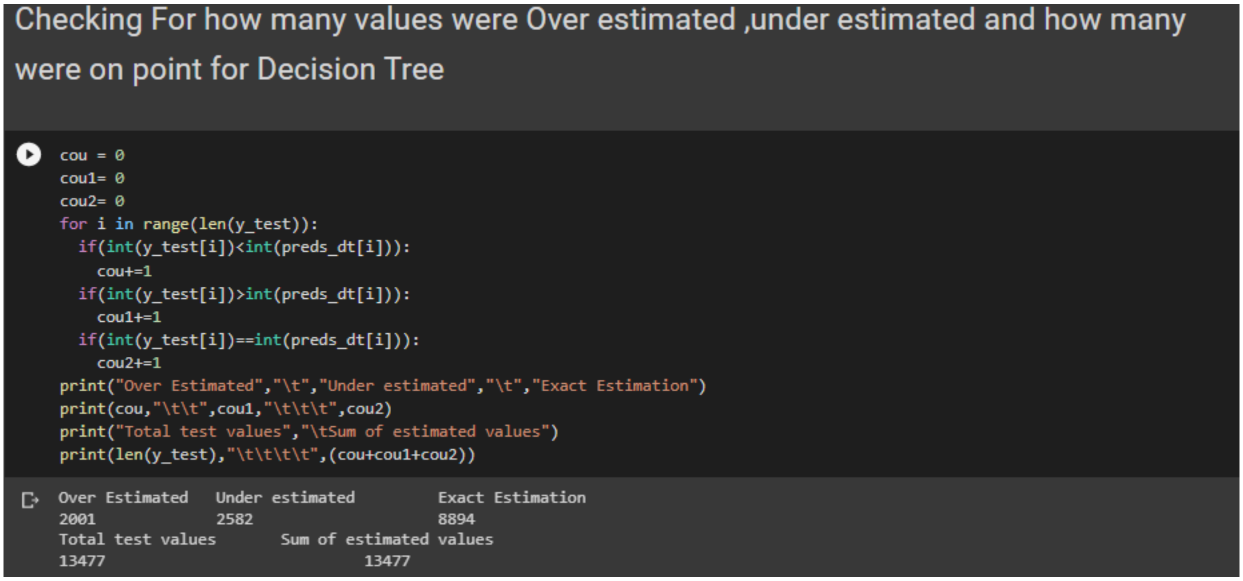
```
Over Estimated    Under estimated          Exact Estimation
7366              6072                     39
Total test values        Sum of estimated values
13477                            13477
```



Accuracy Testing of Decision Tree Model

```
print("R^2:",metrics.r2_score(y_test, preds_dt))
print("Adjusted R^2:",1 - (1-metrics.r2_score(y_test, preds_dt))*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1))
print("MAE:",metrics.mean_absolute_error(y_test, preds_dt))
print("MSE:",metrics.mean_squared_error(y_test, preds_dt))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, preds_dt)))
```

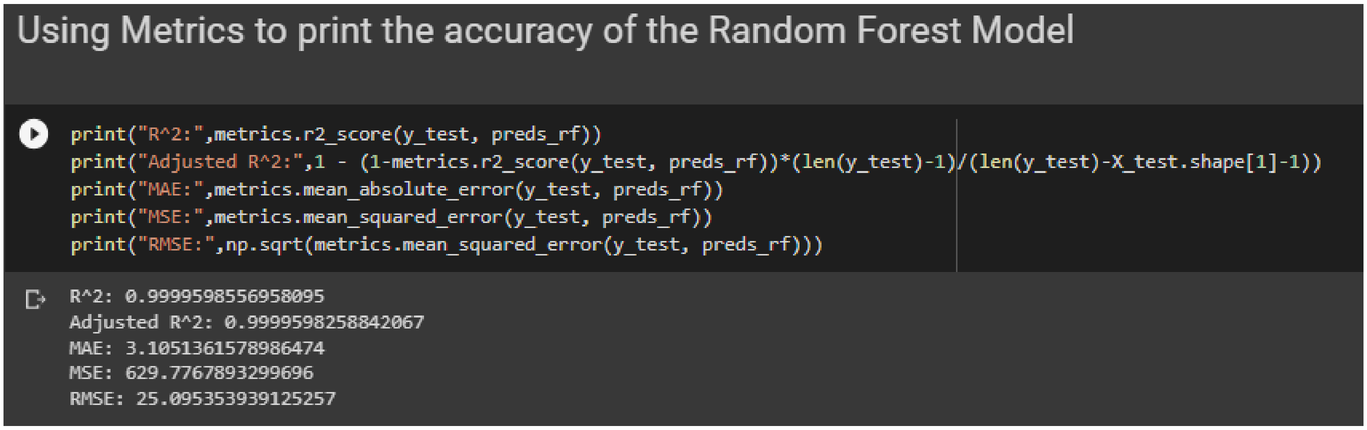
```
R^2: 0.999923806169282
Adjusted R^2: 0.9999237495869037
MAE: 2.854938042591081
MSE: 1195.3154262818134
RMSE: 34.57333403479932
```



Checking For how many values were Over estimated ,under estimated and how many were on point for Decision Tree

```
cou = 0
cou1= 0
cou2= 0
for i in range(len(y_test)):
  if(int(y_test[i])<int(preds_dt[i])):
    cou+=1
  if(int(y_test[i])>int(preds_dt[i])):
    cou1+=1
  if(int(y_test[i])==int(preds_dt[i])):
    cou2+=1
print("Over Estimated","\t","Under estimated","\t","Exact Estimation")
print(cou,"\t\t",cou1,"\t\t\t",cou2)
print("Total test values","\tSum of estimated values")
print(len(y_test),"\t\t\t\t",(cou+cou1+cou2))
```

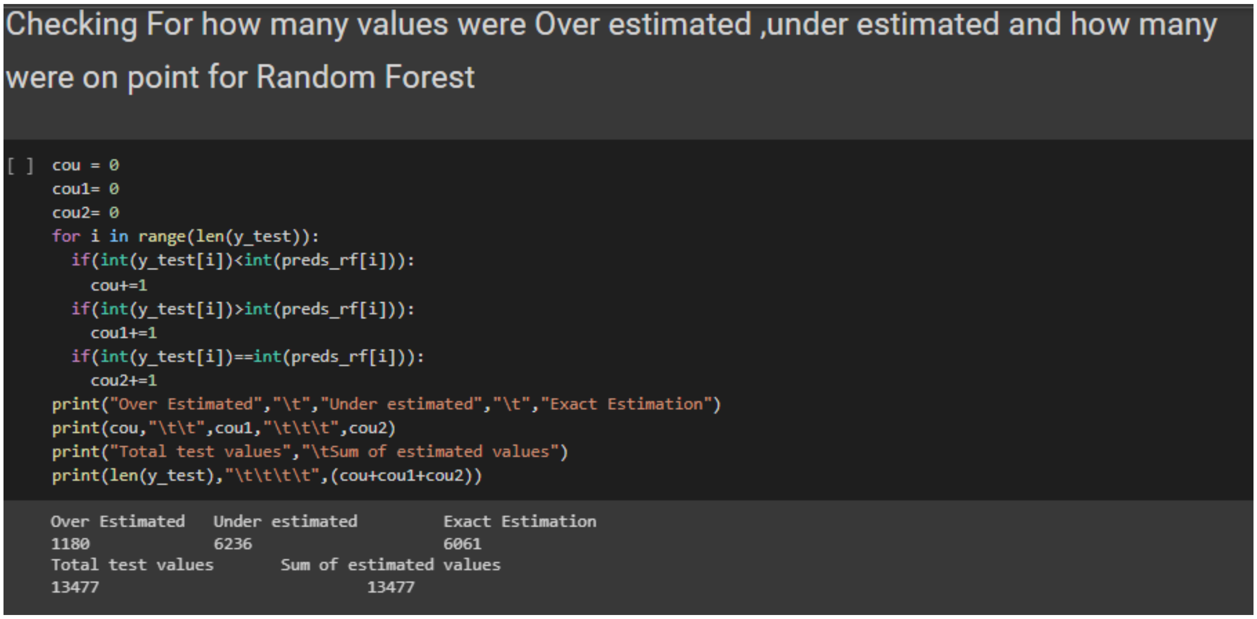
```
Over Estimated    Under estimated          Exact Estimation
2001              2582                     8894
Total test values        Sum of estimated values
13477                            13477
```



Using Metrics to print the accuracy of the Random Forest Model

```
print("R^2:",metrics.r2_score(y_test, preds_rf))
print("Adjusted R^2:",1 - (1-metrics.r2_score(y_test, preds_rf))*(len(y_test)-1)/(len(y_test)-X_test.shape[1]-1))
print("MAE:",metrics.mean_absolute_error(y_test, preds_rf))
print("MSE:",metrics.mean_squared_error(y_test, preds_rf))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, preds_rf)))
```

```
R^2: 0.9999598556958095
Adjusted R^2: 0.9999598258842067
MAE: 3.1051361578986474
MSE: 629.7767893299696
RMSE: 25.095353939125257
```



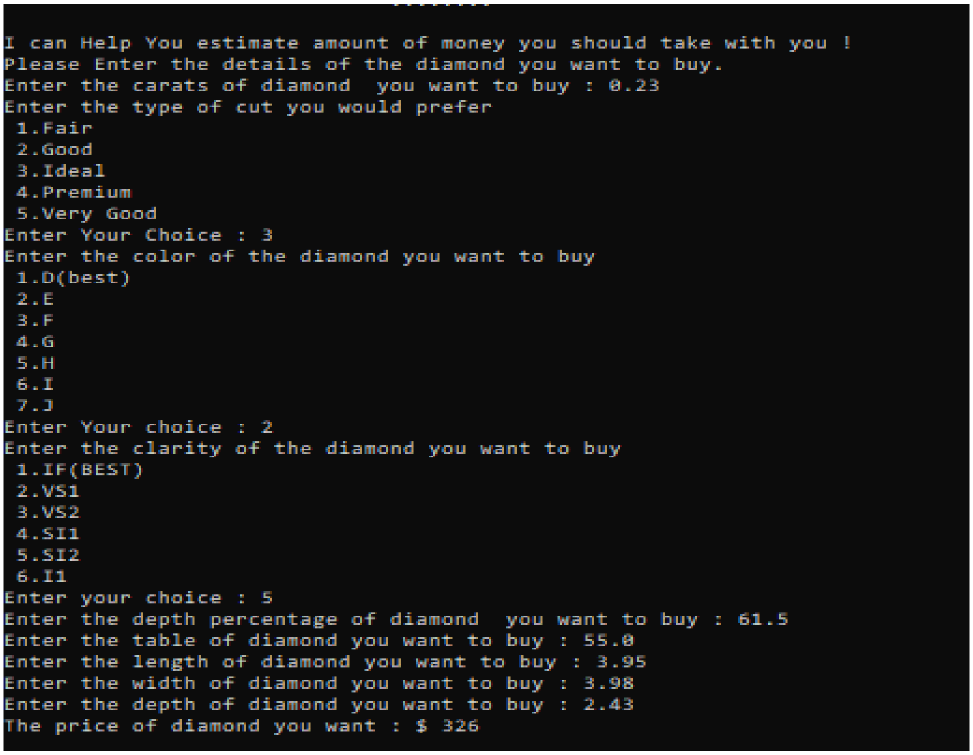
Checking For how many values were Over estimated ,under estimated and how many were on point for Random Forest

```
cou = 0
cou1= 0
cou2= 0
for i in range(len(y_test)):
  if(int(y_test[i])<int(preds_rf[i])):
    cou+=1
  if(int(y_test[i])>int(preds_rf[i])):
    cou1+=1
  if(int(y_test[i])==int(preds_rf[i])):
    cou2+=1
print("Over Estimated","\t","Under estimated","\t","Exact Estimation")
print(cou,"\t\t",cou1,"\t\t\t",cou2)
print("Total test values","\tSum of estimated values")
print(len(y_test),"\t\t\t\t",(cou+cou1+cou2))
```

```
Over Estimated   Under estimated         Exact Estimation
1180             6236                    6061
Total test values        Sum of estimated values
13477                            13477
```

Application Part of the Diamond Prediction Program

```
Hello Welcome To Diamond Genie
```



```
I can Help You estimate amount of money you should take with you !
Please Enter the details of the diamond you want to buy.
Enter the carats of diamond  you want to buy : 0.23
Enter the type of cut you would prefer
 1.Fair
 2.Good
 3.Ideal
 4.Premium
 5.Very Good
Enter Your Choice : 3
Enter the color of the diamond you want to buy
 1.D(best)
 2.E
 3.F
 4.G
 5.H
 6.I
 7.J
Enter Your choice : 2
Enter the clarity of the diamond you want to buy
 1.IF(BEST)
 2.VS1
 3.VS2
 4.SI1
 5.SI2
 6.I1
Enter your choice : 5
Enter the depth percentage of diamond  you want to buy : 61.5
Enter the table of diamond you want to buy : 55.0
Enter the length of diamond you want to buy : 3.95
Enter the width of diamond you want to buy : 3.98
Enter the depth of diamond you want to buy : 2.43
The price of diamond you want : $ 326
```

Discussion

After training different models of machine learning models like linear regression, decision tree, random forest, k-nearest neighbour, and XGBoost we saw that random forest and decision tree were giving the best results even though XGBoost was a more advanced model but it's not working for this model.

But after further analysis, we found that even though the scores of the random forest were very good but decision tree predicts more accurate results than the random forest. Hence we use the decision tree model for predicting the price of diamonds.

Future Scope of the Project

In the future, we can use a deep learning model to see the diamond through the camera and take the measurements and different aspects of a diamond through the camera and then predict the price of the diamond and display the price on the screen.

Conclusion

After training different models of machine learning models like linear regression, decision tree, random forest, k-nearest neighbour, and XGBoost we saw that random forest and decision tree were giving the best results even though XGBoost was a more advanced model but it's not working for this dataset. But after further analysis we found that even though the scores of the random forest were perfect but decision tree predicts more accurate results than the random forest. Hence we use the decision tree model for predicting the price of diamond.This project has future scope as deep learning can be used to predict the price of diamond in real time and this can be used in mines where the diamond is discovered.

References

1. Displayr. (n.d.). What is R-Squared? Displayr. https://www.displayr.com/what-is-r-squared/

2. Allwright, S. (n.d.). How to Interpret MSE. Stephen Allwright. https://stephenallwright.com/interpret-mse/

3. BM Knowledge Center. (n.d.). Adjusted R-squared - IBM Documentation. IBM Knowledge Center. https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-adjusted-r-squared 9-10. Frost, J.(n.d.). How To Interpret R-squared in Regression Analysis - Statistics By Jim.https://statisticsbyjim.com/regression/interpret-r-squared-regression/ 11.Towards Data Science.(n.d.)

4. How To Interpret R-squared in Regression Analysis - Statistics By Jim.https://statisticsbyjim.com/regression/interpret-r-squared-regression/ 11.Towards Data Science.(n.d.)

5. Fürnkranz, J. (n.d.). Decision Tree. SpringerLink. https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8_204

6. Rokach, L., & Maimon, O. (n.d.). Decision Trees. ResearchGate. https://www.researchgate.net/publication/225237661_Decision_Trees

7. Wikipedia contributors. (n.d.). Decision tree. Wikipedia. https://en.wikipedia.org/wiki/Decision_tree

8. Corporate Finance Institute. (n.d.). Decision Tree - Overview, Decision Types, Applications. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/decision-tree/

9. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

10. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System [Computer Science > Machine Learning]. arXiv.org. https://arxiv.org/abs/1603.02754

11. Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (n.d.). Exploratory Data Analysis. ResearchGate. https://www.researchgate.net/publication/308007227_Exploratory_Data_Analysis

12. Wikipedia contributors. (n.d.). Exploratory data analysis. Wikipedia. https://en.wikipedia.org/wiki/Exploratory_data_analysis

13. ResearchGate. (n.d.). Exploratory Data Analysis using Python. ResearchGate. https://www.researchgate.net/publication/341121348_Exploratory_Data_Analysis_using_Python

Exploratory Data Analysis and Price Prediction On Diamond Dataset.

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

Name	**Univ. Roll No.**
Apurva Srivastava	10800119012
Anurag Sharma	10800119018
Sumona Mondal	10800119021
Subhamoy Banerjee	10800119044

UNDER THE GUIDANCE OF

Mrs. SMITA CHAURASIA
Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ASANSOL ENGINEERING COLLEGE
AFFILIATED TO
MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY

May, 2023

Contents

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ASANSOL ENGINEERING COLLEGE

Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305

Certificate of Recommendation

I hereby recommend that the thesis entitled, **"Exploratory Data Analysis and Price Prediction On Diamond Dataset."** carried out under my supervision by the group of students listed below may be accepted in partial fulfilment of the requirement for the degree of "Bachelor of Technology in **Computer Science and Engineering** of Asansol Engineering College under MAULANA ABUL KALAM AZAD UNIVERSITY OF TECHNOLOGY.

Name	**Univ. Roll No.**
Apurva Srivastava	10800119012
Anurag Sharma	10800119018
Sumona Mondal	10800119021
Subhamoy Banerjee	10800119044

…………………………………

(Mrs. Smita Chaurasia)
Thesis Supervisor
Dept. of Computer Science and Engineering,
Asansol Engineering College
Asansol-713305

Countersigned:

………………………………

(Dr. Monish Chatterjee)

Head of the Department
Dept. of Computer Science and Engineering, Asansol
Engineering College,
Asansol-713305

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ASANSOL ENGINEERING COLLEGE

Vivekananda Sarani, Kanyapur, Asansol, West Bengal – 713305

Certificate of Approval

The forgoing thesis is hereby approved as creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project only for the purpose for which it is submitted.

…………………………………

(Mrs. Smita Chaurasia)
Thesis Supervisor
Dept. of Computer Science and Engineering,
Asansol Engineering College,
Asansol-713305

Acknowledgement

It is our great privilege to express our profound and sincere gratitude to our Thesis Supervisor, **Mrs. Smita Chaurasia** for providing us with very cooperative and precious guidance at every stage of the present project work being carried out under his/her supervision. His valuable advice and instructions in carrying out the present study has been a very rewarding and pleasurable experience that has greatly benefited us throughout the course of work.

We would like to convey our sincere gratitude towards **Dr. Monish Chatterjee**, Head of the Department of **Computer Science and Engineering**, Asansol Engineering College for providing us the requisite support for timely completion of our work. We would also like to pay our heartiest thanks and gratitude to all the teachers of the Department of **Computer Science and Engineering**, Asansol Engineering College for various suggestions being provided in attaining success in our work.

We would like to express our earnest thanks to **Mr. Suman Mallick**, of CSE Project Lab for his technical assistance provided during our thesis work.

Finally, I would like to express my deep sense of gratitude to my parents for their constant motivation and support throughout my work.

…………………………………
(Apurva Srivastava)

…………………………………
(Anurag Sharma)

…………………………………
(Sumona Mondal)

…………………………………
(Subhamoy Banerjee)