

***A PROJECT ON***  
**“ML Based Diamond Price Estimator”**

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE COURSE OF  
DIPLOMA IN BIG DATA ANALYSIS



***SUNBEAM INSTITUTE OF INFORMATION TECHNOLOGY***

‘Plot no R/2’, Market yard road, Behind hotel Fulera, Gultekdi Pune – 411037. MH-INDIA

**SUBMITTED BY:**

Atharva Panchakshari (56629)

Siddesh Dagwal (56602)

**UNDER THE GUIDENCE OF:**

Mrs. Pradnya Dindorkar  
Faculty Member, SIIT, PUNE.



## **CERTIFICATE**

This is to certify that the project work under the title ‘ML Based Diamond Price Estimator’ is done by Atharva Panchakshari & Siddesh Dagwal in partial fulfillment of the requirement for award of Diploma in Big Data Analysis Course.

**Mrs. Pradnya Dindorkar**  
**Project Guide**

**Mrs. Pradnya Dindorkar**  
**Course Co-Ordinator**

Date:

## **ACKNOWLEDGEMENT**

A project usually falls short of its expectation unless aided and guided by the right persons at the right time. We avail this opportunity to express our deep sense of gratitude towards Mr. Nitin Kudale (Center Coordinator, SIIT, Pune) and Mrs. Pradnya Dindorikar (Course Coordinator, SIIT, Pune and Project Guide).

We are deeply indebted and grateful to them for their guidance, encouragement and deep concern for our project. Without their critical evaluation and suggestions at every stage of the project, this project could never have reached its present form. Last but not the least we thank the entire faculty and the staff members of Sunbeam Institute of Information Technology, Pune for their support.

Atharva Panchakshari  
DBDA Sept 2021 Batch,  
SIIT Pune

Siddesh Dagwal  
DBDA Sept 2021 Batch,  
SIIT Pune

## Table of Contents

1.	Introduction .....	5
	<b>1.1 Introduction and Objectives:</b> .....	5
	<b>1.2 Dataset Information:</b> .....	5
2.	Problem Definition and Algorithm .....	7
	<b>2.1 Problem Definition:</b> .....	7
	<b>2.2 Algorithm Definition:</b> .....	7
3.	Experimental Evaluation .....	9
	<b>3.1 EDA and Methodology:</b> .....	9
	<b>3.2 Visualization:</b> .....	11
	<b>3.3 Flow Diagram:</b> .....	12
4.	Result and Discussions .....	13
5.	GUI .....	13
6.	Future work And Conclusion .....	14
	<b>6.1 Future Work:</b> .....	14
	<b>6.2 Conclusion:</b> .....	14

# 1. Introduction

## 1.1 Introduction and Objectives:

Demand for Diamonds is increasing day by day and so are the malpractices. In such a competitive market, it is difficult for an average consumer to purchase a diamond at a fair price because there is always a doubt of being duped. Also, a jeweler who outsources diamonds from huge diamond merchants has a fear of getting deceived. This Project is designed to help both the consumer and the businessmen to avoid imprecise valuation preventing deceitful transactions. The objective is simple, when provided with certain features of a diamond we aim at predicting a near accurate price/valuation of that particular diamond which will provide a guideline to the consumer and the industry as a whole.

## 1.2 Dataset Information:

CSV file name: diamonds.csv:

### Feature description:

**Price:** price in US dollars (\$326--\$18,823) This is the target column containing tags for the features.

### The 4 Cs of Diamonds:

**carat (0.2--5.01):** The carat is the diamond's physical weight measured in metric carats. One carat equals 1/5 gram and is subdivided into 100 points. Carat weight is the most objective grade of the 4Cs.

**cut (Fair, Good, Very Good, Premium, Ideal):** In determining the quality of the cut, the diamond grader evaluates the cutter's skill in the fashioning of the diamond. The more precise the diamond is cut, the more captivating the diamond is to the eye.

**color, from J (worst) to D (best):** The color of gem-quality diamonds occurs in many hues. In the range from colorless to light yellow or light brown. Colorless diamonds are the rarest. Other natural colors (blue, red, pink for example) are known as "fancy," and their color grading is different than from white colorless diamonds.

**clarity (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)):** Diamonds can have internal characteristics known as inclusions or external characteristics known as blemishes. Diamonds without inclusions or blemishes are rare; however, most characteristics can only be seen with magnification.

## Dimensions:

- **x length in mm (0--10.74)**
- **y width in mm (0--58.9)**
- **z depth in mm (0--31.8)**
- **depth total depth percentage =  $z / \text{mean}(x, y) = 2 * z / (x + y)$  (43--79):** The depth of the diamond is its height (in mm) measured from the culet (bottom tip) to the table (flat, top surface).
- **table width of the top of the diamond relative to widest point (43--95):** A diamond's table refers to the flat facet of the diamond seen when the stone is face up. The main purpose of a diamond table is to refract entering light rays and allow reflected light rays from within the diamond to meet the observer's eye. The ideal table cut diamond will give the diamond stunning fire and brilliance.

## 2. Problem Definition and Algorithm

### 2.1 Problem Definition:

The dataset has around 9 predictors and 1 predictand. The key is to find right number of features to use as predictors so as to achieve an acceptable level of accuracy of the predictand. We want to fit a model to the training data that is able to predict the price of a diamond as accurately as possible. In fact, our metric of interest will be the Mean Absolute Error and R2 score value. The metric is not very complicated. The further away from the actual outcome our prediction is, the harder it will be punished. Optimally, we exactly predicting the price is of course highly unlikely, but we must try to get as close as possible.

### 2.2 Algorithm Definition:

**Linear regression:** is one of the very basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

**Random forest:** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

**Decision Tree:** algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

**Polynomial regression:** In simple linear regression algorithm only works when the relationship between the data is linear but suppose if we have non-linear data then Linear regression will not capable to draw a best-fit line and It fails in such conditions. consider the below diagram which has a non-linear relationship and you can see the Linear regression results on it, which does not perform well means which do not come close to reality. Hence, we introduce polynomial regression to overcome this problem, which helps identify the curvilinear relationship between independent and dependent variables.

**K-nearest Neighbors Regression:** KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same *neighborhood*. The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error.

**XGBoost:** or extreme gradient boosting is one of the well-known gradient boosting techniques (ensemble) having enhanced performance and speed in tree based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. It is opensource software. Earlier only Python and R packages were built for XGBoost but now it has extended to Java, Scala, Julia and other languages as well.



### 3. Experimental Evaluation

#### 3.1 EDA and Methodology:

Objective of the project is to predict the price of a diamond based on its characteristics to avoid deceitful transaction in the diamond industry. The data set was obtained from 'data.world.com'. The csv file used contains 11 columns. All the columns were studied and it was found that there was a '**Unnamed: 0**' column present which was redundant and hence was dropped.

The columns '**cut**', '**depth**' and '**table**' had some **NaN** values. We filled those values with mode and mean depending upon the type of data it was holding. Data type for column '**z**' was object (string), even though the data was numeric, few strings were removed and column was converted to numeric

Now the categorical data from columns '**cut**', '**clarity**', '**color**' needed to be encoded so as to convert the columns into numeric. This would allow us to use the columns for finding out their co-linearity with our dependent column '**Price**'. For encoding purpose Label Encoder was used, please refer below code:

```
# Create object of LabelEncoder

cut_encoder=LabelEncoder()
color_encoder=LabelEncoder()
clarity_encoder=LabelEncoder()

# Converting object categorical data to numerical categories

df["cut"]=cut_encoder.fit_transform(df["cut"])
df["color"]=color_encoder.fit_transform(df["color"])
df["clarity"]=clarity_encoder.fit_transform(df["clarity"])
```

Now the pre-processed data can be used for checking the relationships and co-linearity so as to decide the independent features (predictors) against the '**price**' (predictand). Using co-relation and trial and error method we concluded that all the columns except '**price**' has to be used as predictor. Removed Impossible Values along with outliers from '**depth**', '**table**', '**clarity**', '**x**', '**y**', '**z**' columns, please refer the below code

```
## Remove all the outliers

df = df[(df["depth"]<75)&(df["depth"]>45)]
df=df[(df["table"]<80) & (df["table"]>40)]
df=df[df["carat"]<3]
df=df[(df["x"]>=2.5) & (df["x"]<=10)]
df=df[(df["y"]>0) & (df["y"]<=20)]
df=df[(df["z"]>=2) & (df["z"]<=6)]
```

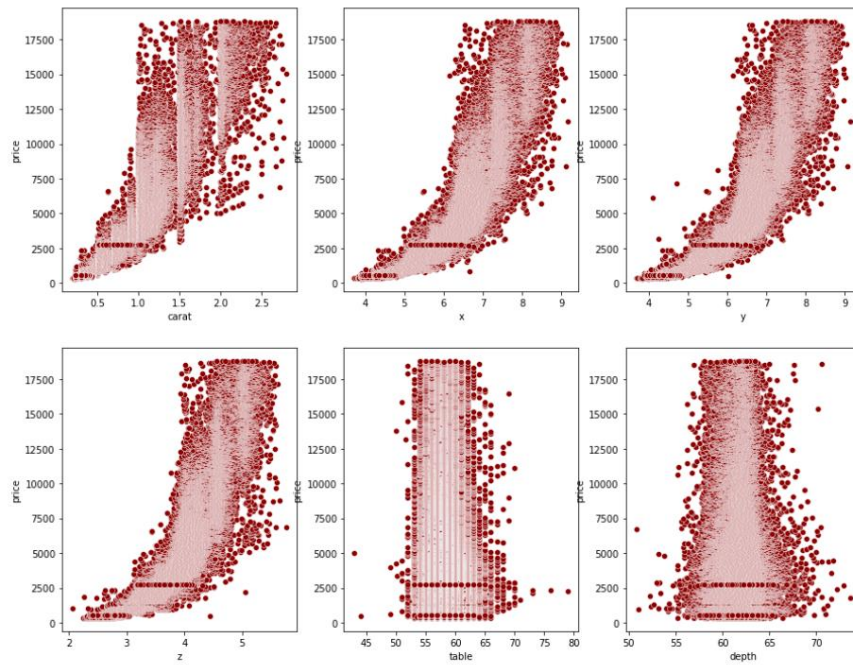


Fig: Co-relation after outlier removal

Above image shows the data from ‘**depth**’, ‘**table**’, ‘**clarity**’, ‘**x**’, ‘**y**’, ‘**z**’ columns after outlier removal. Explored different regression algorithms that can be used for regression model creation. Created multiple pipelines for feeding standardized data to various algorithms like Linear Regression, Random Forest, Decision Tree, XGBoost, Polynomial Regression, K Nearest Neighbor. Trained all above models with Training data set and calculated the accuracies using Test data set and compared all the models. **XGBoost** was found to be the most effective one.

### 3.2 Visualization:

Using the Power BI tool, a plot of average price vs carat has been plotted. Some filters are also provided to analyze the data

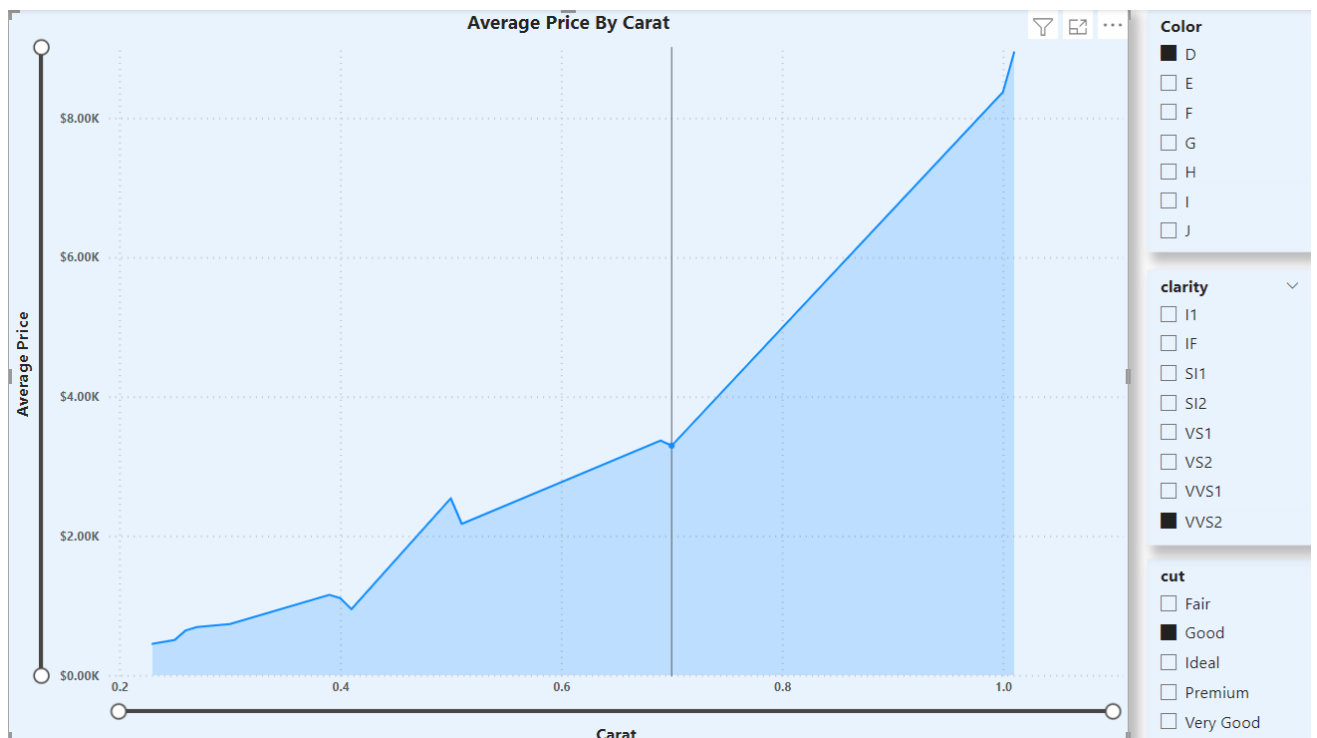


Fig: Average Price vs Carat

Below fig shows the actual data points in relation with the predicted data.

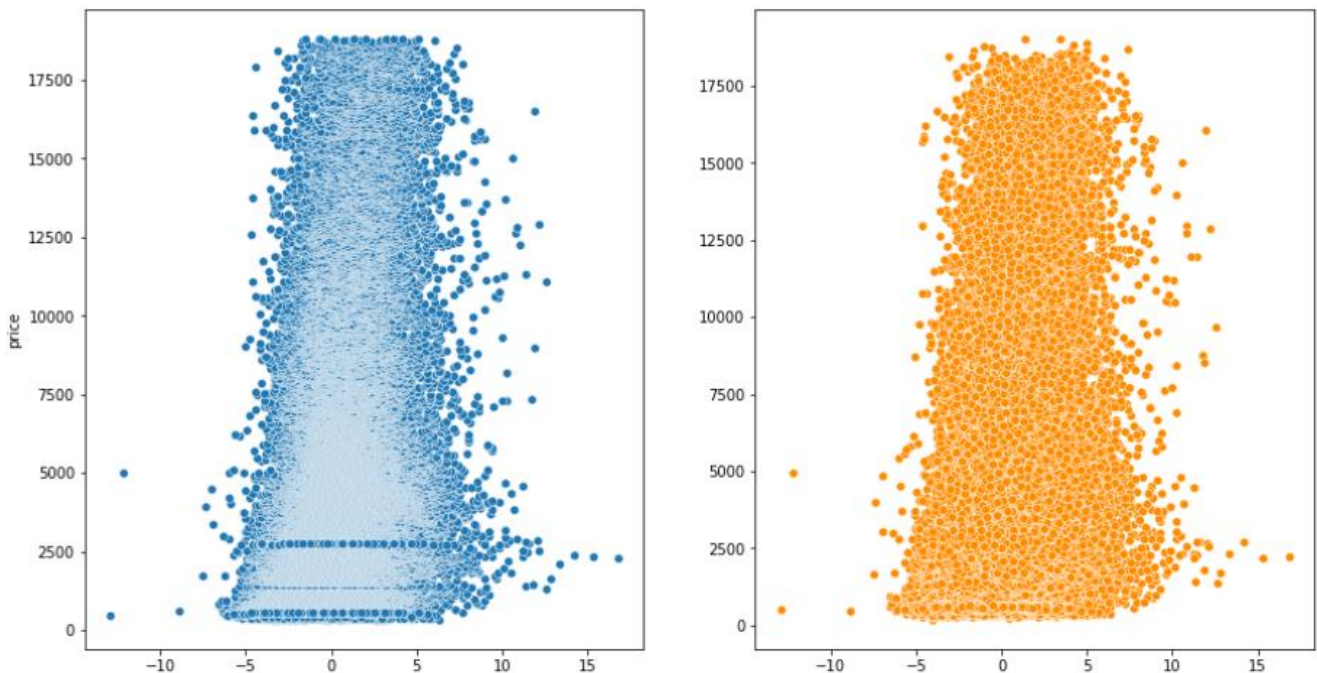
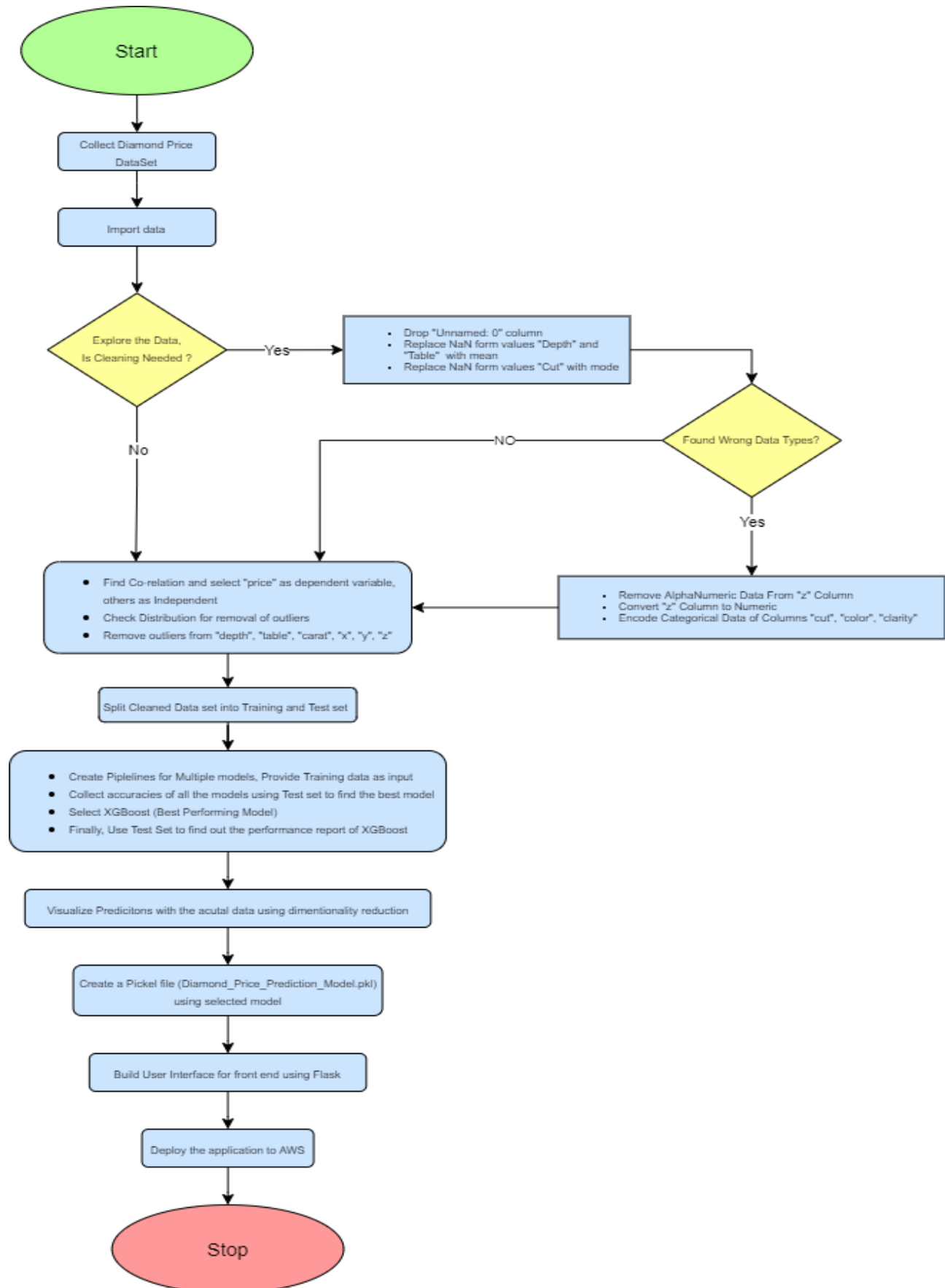


Fig: Comparison of Actual Data and Predicted data

### 3.3 Flow Diagram:



## 4. Result and Discussions

Linear regression, random forest, decision tree, KNN regression, Polynomial regression and XGBoost machine algorithm along with standardization were used to predict the price. Among the given algorithms XGBoost Machine algorithm was the best performing one as it provided the highest R2 score of 0.982.

```
# Based on the accuracies, XGBoost performs the best, Lets use it as the final algorithm
xgb_prediction=pipeline_xgb.predict(x_test)

print("-----Performance Report-----")
print("R Squared Value      :",metrics.r2_score(y_test, xgb_prediction))
print("Adjusted R Squared Value : ",1 - (1-metrics.r2_score(y_test, xgb_prediction))*(len(y_test)-1)/(len(y_test)-x_test.shape[1]))
print("Mean Absolute Error    :",metrics.mean_absolute_error(y_test, xgb_prediction))
print("Mean Squared Error     :",metrics.mean_squared_error(y_test, xgb_prediction))
print("Root Mean Squared Error : ",np.sqrt(metrics.mean_squared_error(y_test, xgb_prediction)))
```

```
-----Performance Report-----
R Squared Value      : 0.9822423824304617
Adjusted R Squared Value : 0.982230492937966
Mean Absolute Error   : 273.2015441522476
Mean Squared Error    : 278097.4573614133
Root Mean Squared Error : 527.3494641709739
```

## 5. GUI

GUI is made using Flask framework. **Flask** is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions. However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools

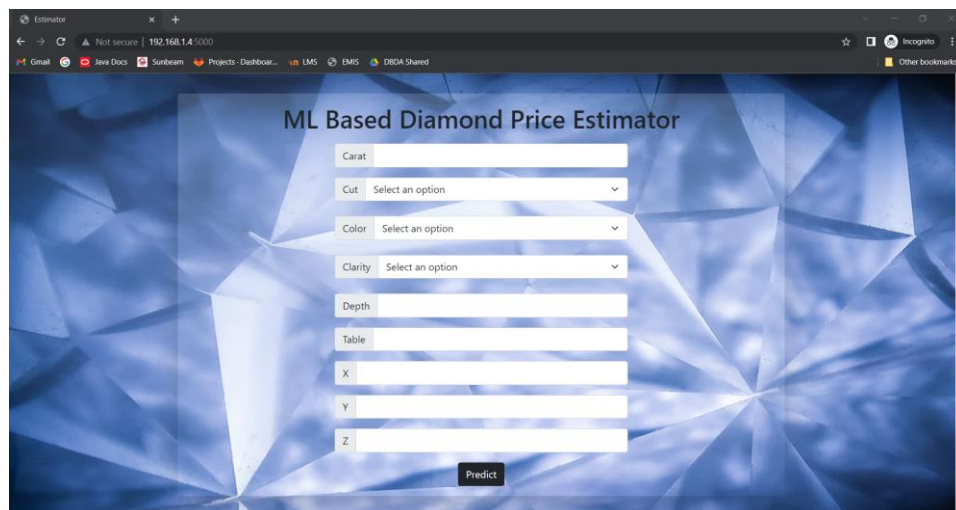


Fig: Frontend of the predictor

## **6. Future work And Conclusion**

### **6.1 Future Work:**

As and when newer data becomes available, the model can be modified for batch learning technique which would further improve the reliability of the model. This will also take into consideration, the effects of time on the prices enhancing credibility of the predictions.

### **6.2 Conclusion:**

- Carat and dimensions (x, y, z) play a major role in deciding the price of a diamond.
- Among the trained models for predicting the price, XGBoost algorithm performs the best.
- Our testing shows predicted values are almost equal to the actual values which makes our model very reliable.
- This project will mitigate the malpractices carried out in the industry at least to some extent.