School of Engineering

Report On

"Diamond Price Prediction System"

Student's Full Name	KARTAVYA JADAV, VASU GOLAKIYA, KISHAN MANGUKIYA,
Enrollment No.	20SE02ML020, 20SE02ML017, 20SE02ML028
Branch:	B. TECH AIML SEM (5)

Supervised by:
Ms. Zinal Gohil

P P SAVANI
U N I V E R S I T Y

CERTIFICATE

This is to certify that Mr. KARTAVYA JADAV,

Enrollment No. 20SE02ML020 from the Department of

B. Tech in Artificial Intelligence and Machine Learning has successfully completed the

Minor Project / Major Project from 19/07/2022 to 06/10/2022.

Date:

Signature: Stamp / Seal

P P SAVANI
UNIVERSITY

CERTIFICATE

This is to certify that Mr. VASU GOLAKIYA,

Enrollment No. 20SE02ML017 from the Department of

B. Tech in Artificial Intelligence and Machine Learning has successfully completed the

Minor Project / Major Project from 19/07/2022 to 06/10/2022.

Date:

Signature: Stamp / Seal

CERTIFICATE

This is to certify that Mr. KISHAN MANGUKIYA,

Enrollment No. 20SE02ML028 from the Department of

B. Tech in Artificial Intelligence and Machine Learning has successfully completed the

Minor Project / Major Project from 19/07/2022 to 06/10/2022.

Date:

Signature: Stamp / Seal

<u>Acknowledgement</u>

A study or a project of this volume can never be the outcome of a single person or just a mere group of dedicated students, so we express our profound sense of gratitude to those who extended their wholehearted help and support to us in completing our project because successful completion of any work requires guidance and help from several people.

Firstly, it gives us immense pleasure to acknowledge our institute **PP SAVANI UNIVERSITY** for providing us an opportunity in developing a project on **Diamond Price Prediction System.**

In addition, we wish to express our deep sense of gratitude to **Ms. Zinal Gohil**, for permitting us to carry out this project work and for her guidance and support.

We give sincere thanks to **Ms. Zinal Gohil,** for her special concern and providing sufficient information related to the topic, which helped us in completing the project work in time and her timely guidance had been a source of inspiration in the conduct of our project work. Last but not the least, we extend our whole-hearted gratitude for the invaluable contribution of our parents for their blessings and earnest affection and all those persons 'behind the veil' for their necessary support, which enabled us to complete this project.

Table of Contents

1. Introduction

In this emerging world of computers, almost all-manual system has switched to automated and computerized system. Therefore, we are developing the software for "Diamond Price Prediction System" to model the present system and to remove the drawbacks of the present system. This project explores how computer technology can be used to solve the problem of user.

This being a big step in terms of improvement in the Diamond companies it is widely accepted across the country. Rather than designing manually, we have made use of computer. Use of computer has solved many problems, which are faced during manual calculation. Once data are fed, it can perform accurate functions. Therefore, to reduce the complexity and efficiency a versatile and an outsourcing price prediction system has been developed.

This project introduces Diamond price prediction system. This project is developed in Php, HTML, CSS, Python, JavaScript languages. All most all the header files have been used in this project. Proper comments have been given at desired locations to

make the project user friendly. Various functions and structures are used to make a complete use of this language.

The companies the following specification to predict the price of the Diamond:

- price: The price of the Diamond
- carat: The carat value of the Diamond
- cut: The cut type of the Diamond, it determines the shine
- color: The color value of the Diamond
- clarity: The carat type of the Diamond
- depth: The depth value of the Diamond
- table: Flat facet on its surface — the large, flat surface facet that you can see when you look at the diamond from above.
- x: Width of the diamond
- y: Length of the diamond
- z: Height of the diamond

After getting these specifications they just must enter it in our trained system then the most accurate price of the Diamond (according to the specifications) will be given the system.

This project is dedicated to model the existing Diamond price prediction system that aims at development of Diamond Price Prediction System that facilitates the Diamond companies to

manage their gains and loss in more efficient, and accurate manner.

And now one can easily plan to buy the diamond as they will know the dealer from whom he/she is buying is giving him/her in good price or not. And moreover, the Diamond companies will also be able to get accurate price and increase their gains.

1.1 Objective

Our project Diamond Price Prediction system with an objective to predict the price of Diamond in more efficient, easier, and fast way.

This project explores how computer technology can be used to solve the problem of user. The main objectives provided by this software are as follows:

- ➢ As prediction of the price will be done by trained machine there will be less chance of mistake.
- ➢ Loss of Diamond companies will decrease.
- ➢ It can also be used by people to predict the price of their future diamond (which he/she is going to buy).

1.2 Research Questions

Research question: Which machine learning algorithm performs better and has the most accurate result in Diamond price prediction?

1.3 Limitations

The request contains a list of features, that matches the public dataset's features, that is desired to

be available when the data is sent. There is no guarantee that the data will be available in time nor contains the exact requested list of features. Thus, there might be a risk that the access will be denied or delayed. If so, the study will be accomplished based only on the public dataset.

Moreover, this study will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this study.

2. Background

2.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is a supervised technique used to estimate the relationship between one dependent variable and more than one independent variables. Identifying the correlation and its cause-effect helps to make predictions by using these relations [4]. To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Multiple Linear Regression is prone to many problems such as multicollinearity, noises, and overfitting, which effect on the prediction accuracy.

Regularized regression plays a significant part in Multiple Linear Regression because it helps to reduce variance at the cost of introducing some bias, avoid the overfitting problem and solve ordinary least squares (OLS) problems. There are two types of regularization techniques L1 norm (least absolute deviations) and L2 norm (least squares). L1 and L2 have different cost functions regarding model complexity.

2.2 Decision Tree

Decision Trees are used in classification and regression tasks, where the model (tree) is formed of nodes and branches. The tree starts with a root node, while the internal nodes correspond to an input attribute. The nodes that do not have children are called leaves, where each leaf performs the prediction of the output variable.
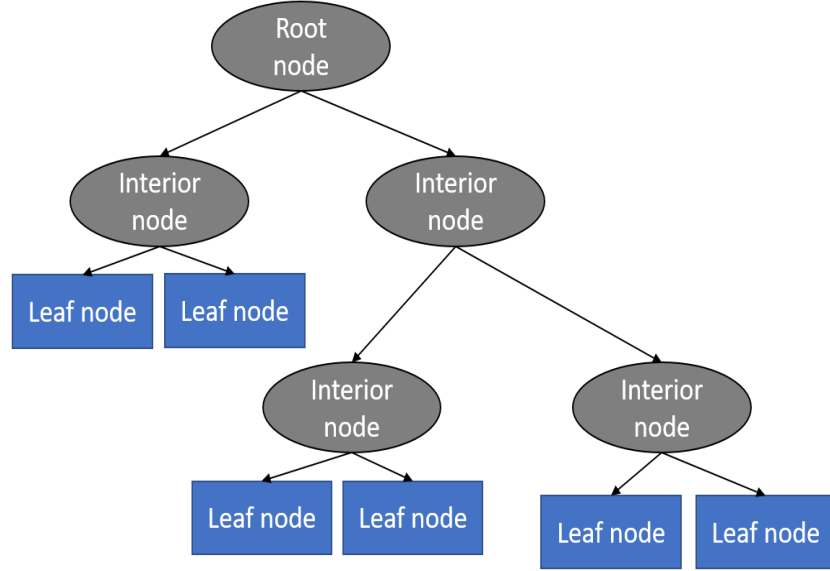
Figure 1. Decision Tree

A Decision Tree can be defined as a model:

$$\varphi = X \longmapsto Y$$

Where any node t represents a subspace $X_t \subseteq X$ of the input space and internal nodes t are labelled with a split s_t taken from a set of questions Q. However, to determine the best separation in Decision Trees, the Impurity equation of dividing the nodes should be taken into consideration, which is defined as:

$$\Delta i\,(s, t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Where $s \in Q$, t_L and t_R are left and right nodes, respectively. pL and pR are the proportion $^N{}_N\underline{tL}t$ and $^N N\underline{tR}t$ respectively of learning samples from $\mathcal{L}_t$ going to t_L and t_R respectively. N_t is the size of the subset $\mathcal{L}_t$.

2.3 Random Forest Regression

A Random Forest is an ensemble technique qualified for performing classification and regression tasks with the help of multiple decision trees and a method called Bootstrap Aggregation known as Bagging. Random Forest is a model that constructs an ensemble predictor by averaging over a collection of decision trees. Therefore, it is called a forest, and there are two reasons for calling it random. The first reason is growing trees with a random independent bootstrap sample of the data. The second reason is splitting the nodes with arbitrary subsets of features. However, using the bootstrapped sample and considering only a subset of the variables at each step results in a wide variety of trees. The variety is what makes Random Forest more effective than individual Decision Tree.
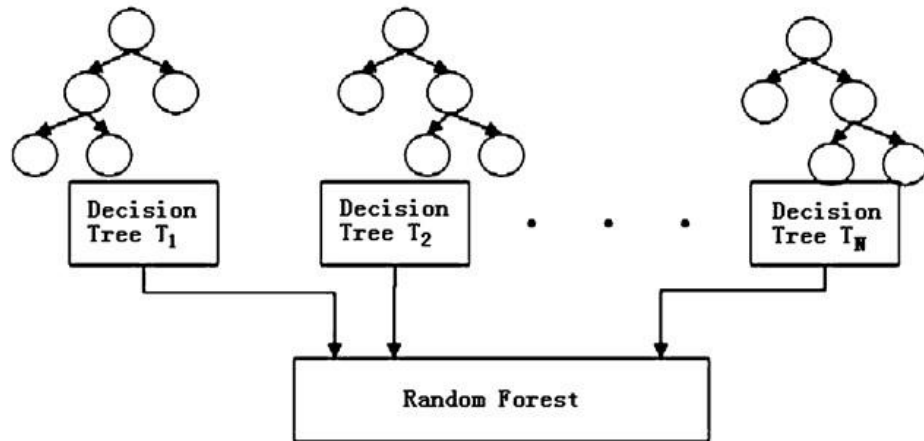

Figure 2. Random Forests

To improve prediction performance, Random Forest acquires out-of-bag (OOB) estimates, which is based on the fact that, for every tree, approximately $e^{1} \approx 0.367 \; or \; 36\%$ of cases are not in the bootstrap sample. There are several advantages to using OOB. One advantage

is that the complete original example is used both for constructing the Random Forest classifier and for error estimation. Another advantage is its computational speed, especially when dealing with large data dimensions.

2.4 KNN 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.

While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.
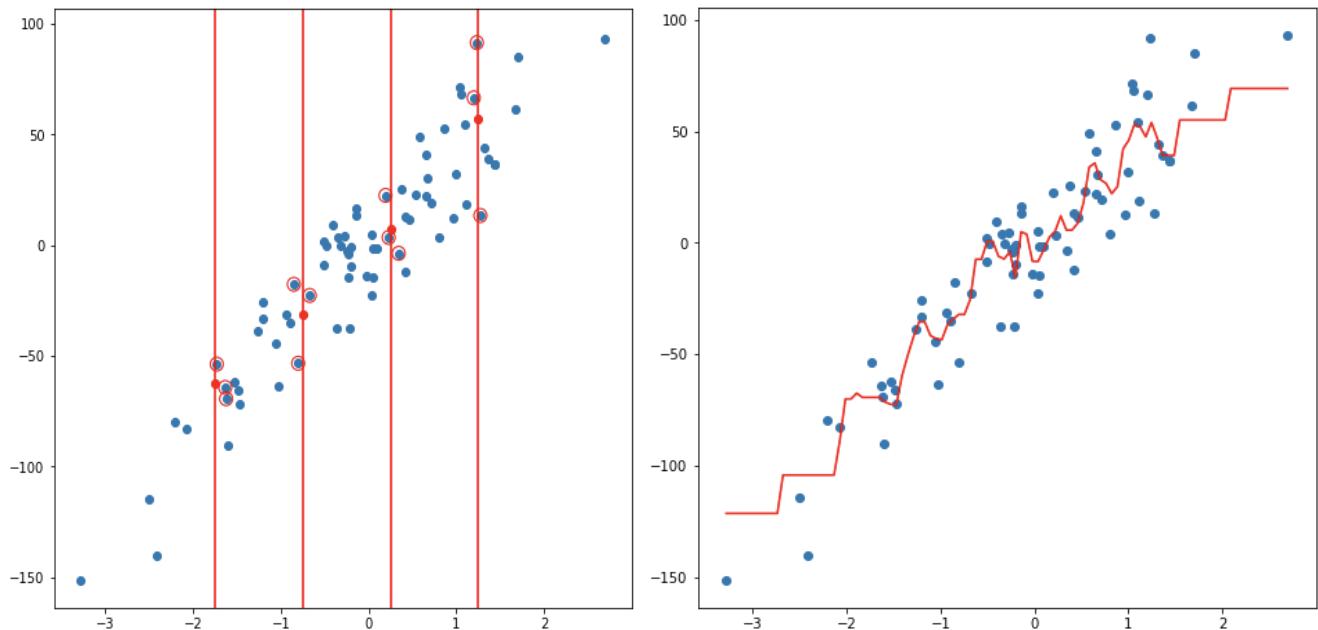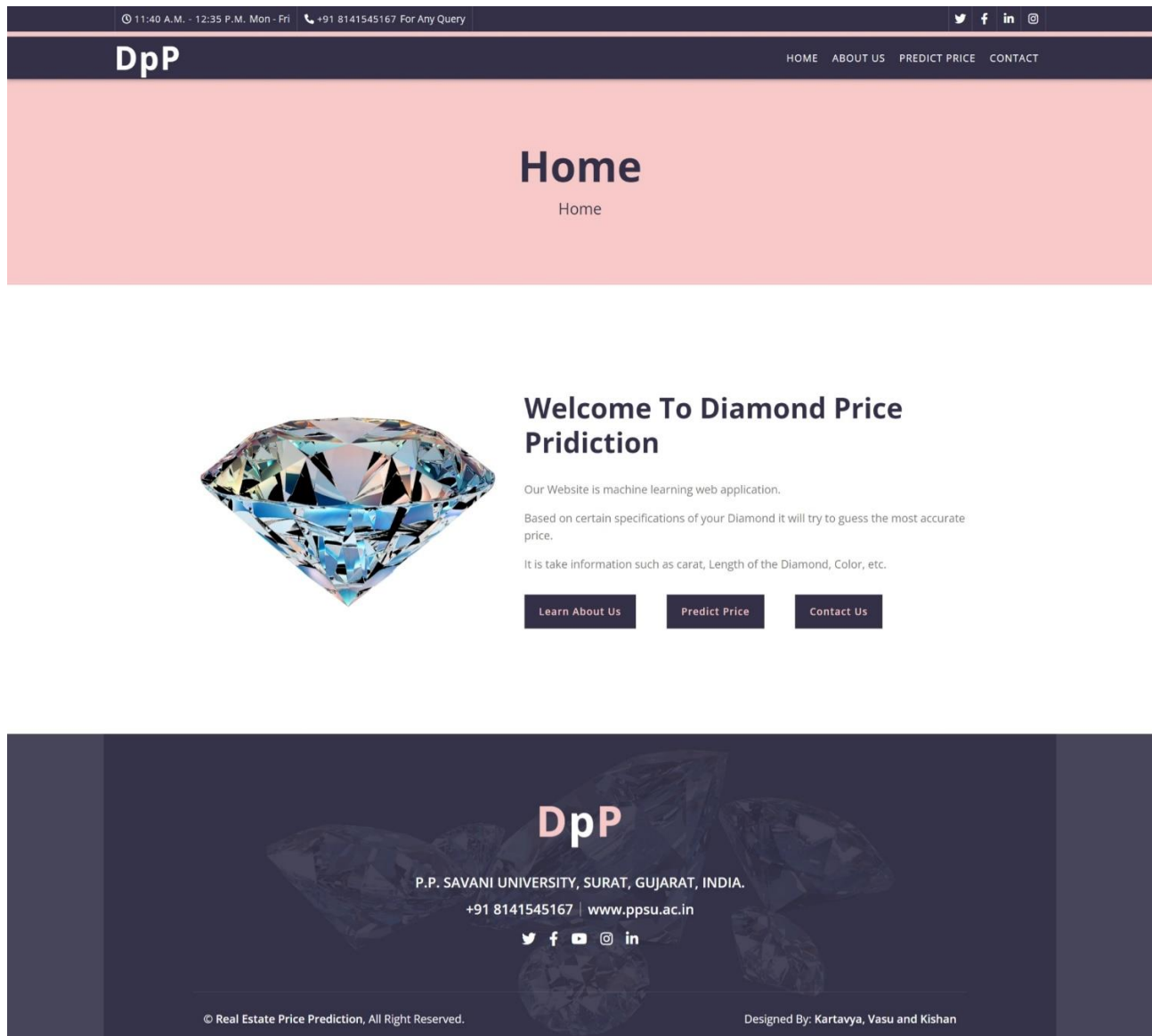

Figure 3. KNN Regression

3. Method

This study has been organized through theoretical research and practical implementation of regression algorithms. We also created a web application using HTML, CSS, JavaScript, and Flask for using the model to predict the price of the Diamond.

3.1 Screenshots of Web Application
3.1.1 Home Page

3.1.2 About Us Page

DpP

HOME   ABOUT US   PREDICT PRICE   CONTACT

About Us

Home / About Us

—— Learn About Us ——

Welcome to Diamond Price Prediction

We are a team of college students working on this project like it's our full time job. Any amount would help support and continue development on this project and is greatly appreciated.

Learn More

—— Project Developers ——

Our Expert Developers

Kartavya Jadav   Vasu Golakiya   Kishan Mangukiya

DpP

P.P. SAVANI UNIVERSITY, SURAT, GUJARAT, INDIA.

+91 8141545167 | www.ppsu.ac.in

© Real Estate Price Prediction, All Right Reserved.   Designed By: **Kartavya, Vasu and Kishan**

3.1.3 Price Predication Page

DpP

HOME   ABOUT US   PREDICT PRICE   CONTACT

Price

Home / Price

Description

Description about the Values needed to enter to predict the price of the Diamond.

1. carat: The carat value of the Diamond

2. y: Length of the Diamond

3. color: The color value of the Diamond

For "J" : 1,

For "I" : 2,

For "H" : 3,

For "G" : 4,

For "F" : 5,

For "E" : 6,

For "D" : 7

4. clarity: The carat type of the Diamond

For "I1" : 1,

For "SI1" : 2,

For "SI2" : 3,

For "VS1" : 4,

For "VS2" : 5,

For "VVS1" : 6,

For "VVS2" : 7

For "IF" : 8

Enter specifications to predict the price

carat

y

color

PLEASE ENTER NUMBERS FROM 1 TO 7 ONLY

clarity

PLEASE ENTER NUMBERS FROM 1 TO 8 ONLY

Predict Price

DpP

P.P. SAVANI UNIVERSITY, SURAT, GUJARAT, INDIA.

+91 8141545167 | www.ppsu.ac.in

17

3.1.4 Contact Us Page

3.2 Experiment

The experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. These stages can be defined as:

- ➢ Pre-processing: both datasets will be checked and pre-processed using the methods from section 4.2. These methods have various ways of handling data. Thus, the preprocessing is done on multiple iterations where each time the accuracy will be evaluated with the used combination.

- ➢ Data splitting: dividing the dataset into two parts is essential to train the model with one and use the other in the evaluation. The dataset will be split 70% for training and 30% for testing.

- ➢ Evaluation: the accuracy of both datasets will be evaluated by measuring the R2 and RMSE rate when training the model alongside an evaluation of the actual prices on the test dataset with the prices that are being predicted by the model.

- ➢ Performance: alongside the evaluation metrics, the required time to train the model will be measured to show the algorithm vary in terms of time.

- ➢ Correlation: correlation between the available features and Diamond price will be evaluated using the Pearson Coefficient

Correlation to identify whether the features have a negative, positive or zero correlation with the Diamond price.

3.2.1 Evaluation Metrics

The prediction accuracy will be evaluated by measuring the R-Squared (R2), and Root Mean Square Error (RSME) of the model used in training. R2 will show if the model is overfitted, whereas RSME shows the error percentage between the actual and predicted data, which in this case, the diamond prices.

3.2.2 Computer Specifications

The needed time to train the model depends on the capability of the used system during the experiment. Some libraries use GPU resources over the CPU to take a shorter time to train a model.

Table 1. Computer Specifications

Operating System	Windows 11
Processor	Intel 11th gen core-i7-11800H @2.30GHz
RAM	16 GB
Graphics card	NVIDIA RTX 3070
SSD	1TB

3.2.3 Algorithm's Properties/Design

The algorithms used in this study have different properties that will be used during the implementation. The experiment is done with the IDE Spyder using Python as a programming language. However, in all algorithms, the data is split into four variables, namely, X_train,

X_test, y_train, and y_test, by using train_test_split class from the library sklearn.model_selection. In addition, in all algorithms, the train_test_split class takes as parameters the independent variables, which is the data, the dependent variable, which is the SalePrice, test_size = 0.30, and random_state = 42. The properties and design of each algorithm are as below:

Multiple linear:

Multiple linear is implemented using the LinearRegression from the library sklearn.linear_model. This library takes only the independent variables and dependent variable as parameters.

Decision Tree:

Decision Tree is implemented using the Tree from the library sklearn.tree. This library takes only the independent variables and dependent variable as parameters.

Random Forest:

Randomforest is implemented using the sklearn.ensemble.RandomForestRegressor library. This library takes several parameters to set up the model properties. The model consists of 1200 tree where the max depth of the tree is set to 60.

KNN Regression:

KNN Regression is implemented using the sklearn.neighbors kneighborsregressor library. This library takes only the number nearest neighbors as parameter.

4. Conclusion

4.1 Research Question Result

Research question: Which machine learning algorithm performs better and has the most accurate result in Diamond price prediction?

Research answer: KNN Regression made the best performance overall when accuracy taking into consideration. It has achieved the best performance due to its nearest neighbor technique.

5. Bibliography

- www.python.org/doc\
- www.codewithharry.com
- www.scribd.com