iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52741

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Diamond Price Prediction Using Machine Learning Algorithms

Prof. Amruta. A. Mankawade[1], Chinmay Kokate[2], Koustubh Soman[3], Atharva Mohite[4], Ayush Vispute[5], Omkar More[6]
[1, 2, 3, 4, 5, 6]*Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune*

Abstract: ***Diamonds are one of the most powerful, precious and treasured naturally occurring materials that form from pre-historic carbon. Nevertheless, as opposed to gold and silver, diamond pricing is very complicated as there are many features to consider when determining the price. The aim of this paper remains to put forth a more optimal solution for the purpose of diamond price projection. This involves training a distinct machine learning model on the available diamond dataset to predict diamond prices established by considering several attributes using algorithms such as linear regression, decision trees, and K-nearest neighbours. In addition, a correlative survey of different machine learning regression models is executed for the purpose of diamond price prediction.***

Keywords: Machine Learning Algorithms, Diamond Price Prediction, Linear Regression, K NN, Decision Trees, Flask.

I. INTRODUCTION

Diamonds prevail to be one of the sparsely scattered elements and most sought after naturally occurring minerals, pertaining from carbon. It remains to be the hardest material currently in existence. It flaunts a high thermal conductivity and excellent chemical resistance. They can also be considered as gemstones. It is one of the more expensive gemstones than any other gemstone combined. Diamonds have exceptional optical properties which makes them increasingly useful for multispectral optical applications. Moreover, diamonds robustness, habits, fashion, and paramount marketing from diamond jewellers. A combination of all other gems. Diamonds are growing in popularity due to their optical properties. Diamonds exhibit a luster of a certain refractive index which gives them a term known as 'adamantine'. This luster property reflects an excess percentage of the light that hits the diamond's plane. This is the property that gives the rough diamond its “brilliance”. Diamonds are very hard materials in nature due to which they are majorly used for their abrasiveness. This provides numerous industrial uses for them. Tools like saw blades, grinding wheels and drills are ingrained with tiny diamond stones. The main motivation for this paper is to initiate a supervised machine learning technique for predicting diamond prices (expressed in US dollars ($)). Determine exact results using Kaggle's diamond dataset and supervised machine learning techniques. We also compare linear regression, decision tree, and KNN results to determine which is better suited for the task.

II. LITERATURE REVIEW

Many studies have attempted to predict diamond prices using various techniques. For example, José [2] employed data mining techniques such as M5P, linear regression, and neural networks, with the M5P model showing a high level of accuracy. Singfat [3] used multiple linear regression (MLR) for examining the relationships between diamond prices and the 4Cs (carat weight, cut, colour, and clarity). MLR is a common and appropriate data mining model for diamond datasets. While several machine learning algorithms have been used to predict diamond and gold prices, the challenge resides in selecting the ideal model using pre-processing and correlation approaches [4]. Typically, diamond prices are expressed in US dollars, but the relationship between price and carat weight is not always linear. This is because heavier diamonds are generally more expensive than lighter ones, and the trend of a high correlation between carat weight and price seems to be fading [5]. A scatter plot visualization of the Kaggle diamond dataset can help better understand this relationship and the volatility of diamond prices for heavier stones.

III. METHODOLOGY/EXPERIMENTAL

The Methodology section of this project is divided into three subparts, Sections A, B, and C. In Section A, the tool used for data analysis is discussed. Section B covers data acquisition, and in Section C, the statistical evaluations that can be made are described [5]. The dataset is analysed using supervised learning techniques, which provides a formidable mechanism for processing and classifying data using machine learning algorithms. The aim of supervised learning is that previously labelled data is used for the learning algorithm; which can then be made use of as a base to predict the classification of some more unlabelled data using machine learning algorithms [7].

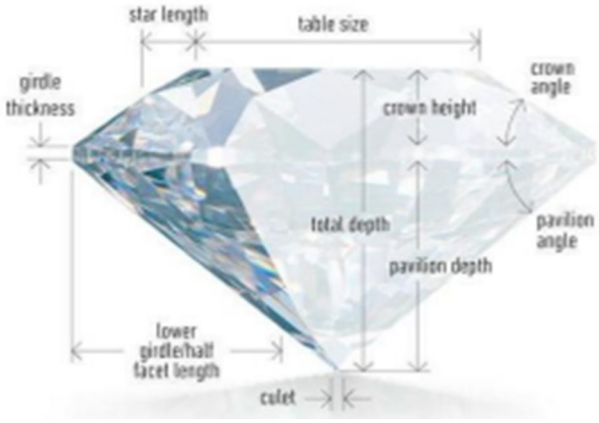
Linear regression is used to ascertain the scope of the linear relationship between dependent and independent variables. Decision trees in this case are made use of to identify the approach most likely to achieve our target goal. K-NN is a non-parametric, supervised learning classifier that uses proximity to make predictions about the grouping of individual data points.
The Flask framework is used for GUI development. Flask is a lightweight Python web framework that provides useful tools and features for creating web applications in the Python language. It is an accessible framework for new developers and allows for the quick creation of web applications using a single Python file.

A. Tools Used

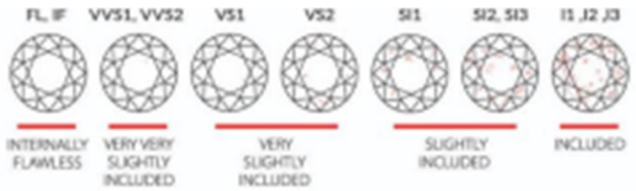
Python 3 is used for analysing the dataset. It is an open-source, high-level programming language that is highly useful for object-oriented programming. Python is majorly used to research and by scientific factions due to its easy-to-use and understandable language syntax. HTML is used to create the website structure.

B. Data Acquisition

Kaggle is an online platform that brings together data scientists and machine learning enthusiasts from around the world. It offers a collaborative environment where participants can access datasets, tools, and resources to solve real-world problems through data-driven competitions. With a focus on innovation and community engagement, Kaggle encourages ethical practices, original work, and knowledge sharing. By fostering collaboration and providing a platform for learning, Kaggle has become a hub for data science enthusiasts to showcase their skills, develop models, and contribute to the broader field of artificial intelligence. The diamond dataset from Kaggle provides the necessary features for the Diamond Dataset analysis.



(Diamond cut parameters) Fig. 1



(Diamond Clarity chart) Fig. 2

(Diamond Colour chart) Fig. 3

C. Evaluating the Statistics

The first step in evaluating the statistics is to split the dataset into a training set (70%) and a test set (30%). The test set allows the model to make predictions on values that it has never seen before. However, taking random samples from the dataset can introduce significant sampling bias.

Therefore, to avoid this bias, the data will be divided into different homogeneous subgroups called strata. This is known as stratified sampling. Since carat is the most important parameter to predict the price of diamonds, we will use it for stratified sampling.
To manage large errors in this large dataset, we will use root mean square error (RMSE) instead of mean absolute error. RMSE is most useful for particularly undesirable large errors [14]. We will first find the mean squared error (MSE) to check the performance. The function will plot a graph to show how well our algorithm has predicted the data. The results obtained from each evaluation will be averaged together to compute a final score. After finding the final score, the final model will be fit on the entire dataset to begin the processing [15].

IV. RESULTS AND DISCUSSIONS

After conducting various experiments and analysing the results, it can be concluded that the supervised learning methods like linear regression, decision tree, and KNN can effectively be used to evaluate diamond prices. The Decision Tree Regressor algorithm showed the best performance, with an accuracy of around 87.49% to 88%. In future work, it would be useful to incorporate unsupervised models to further enhance the accuracy and robustness of diamond price predictions using the dataset.

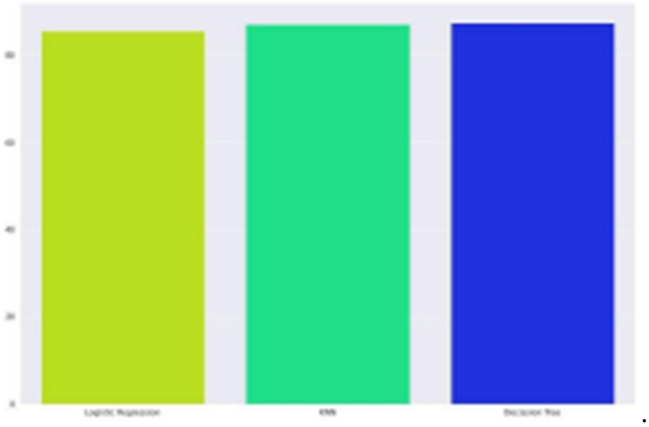


Fig. 4 Accuracies of Models

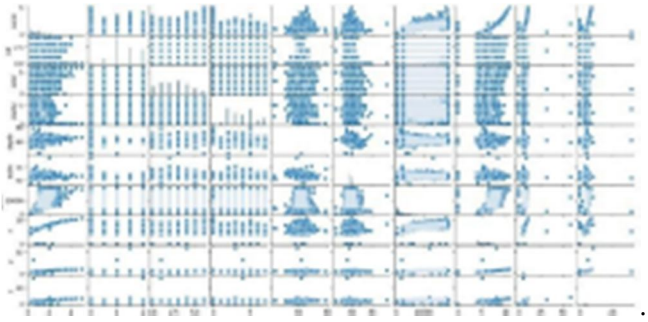


Fig. 5

Fig. 6 Confusion Matrix

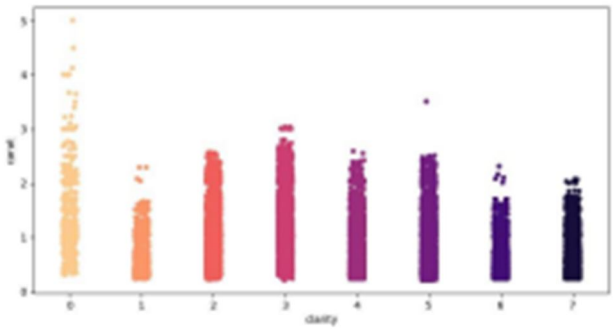


Fig. 7 Clarity

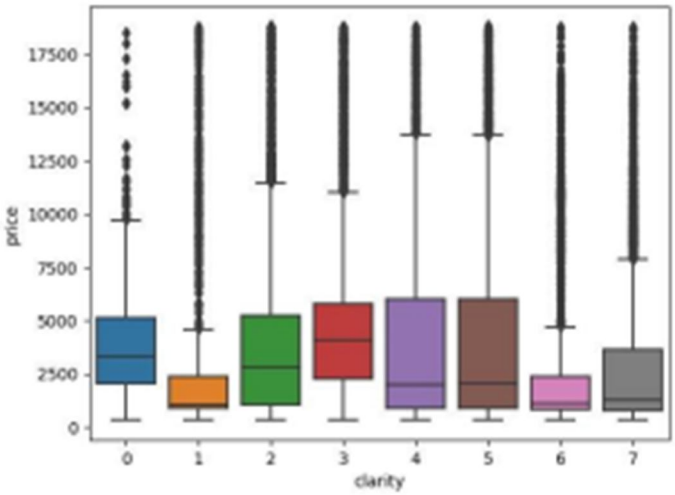


Fig. 8 Clarity

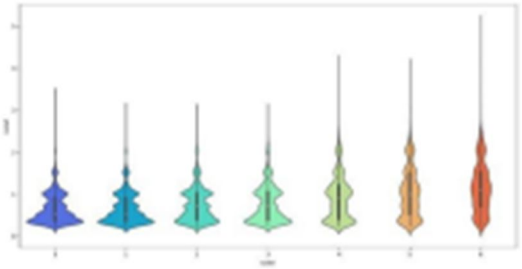


Fig. 9 Color

V. CONCLUSION

Diamond price prediction using machine learning algorithms is a complex task that involves analysing various factors that affect the value of diamonds, such as carat weight, colour, clarity, cut, and other market trends. Machine learning algorithms are a promising approach to predict diamond prices as they can identify patterns and relationships between the various factors that affect the value of diamonds.

The prediction of diamond prices using machine learning algorithms can be achieved through supervised learning techniques such as regression, decision trees, and neural networks. These algorithms can use historical data to predict the future prices of diamonds with high accuracy, enabling buyers and sellers to make informed decisions.

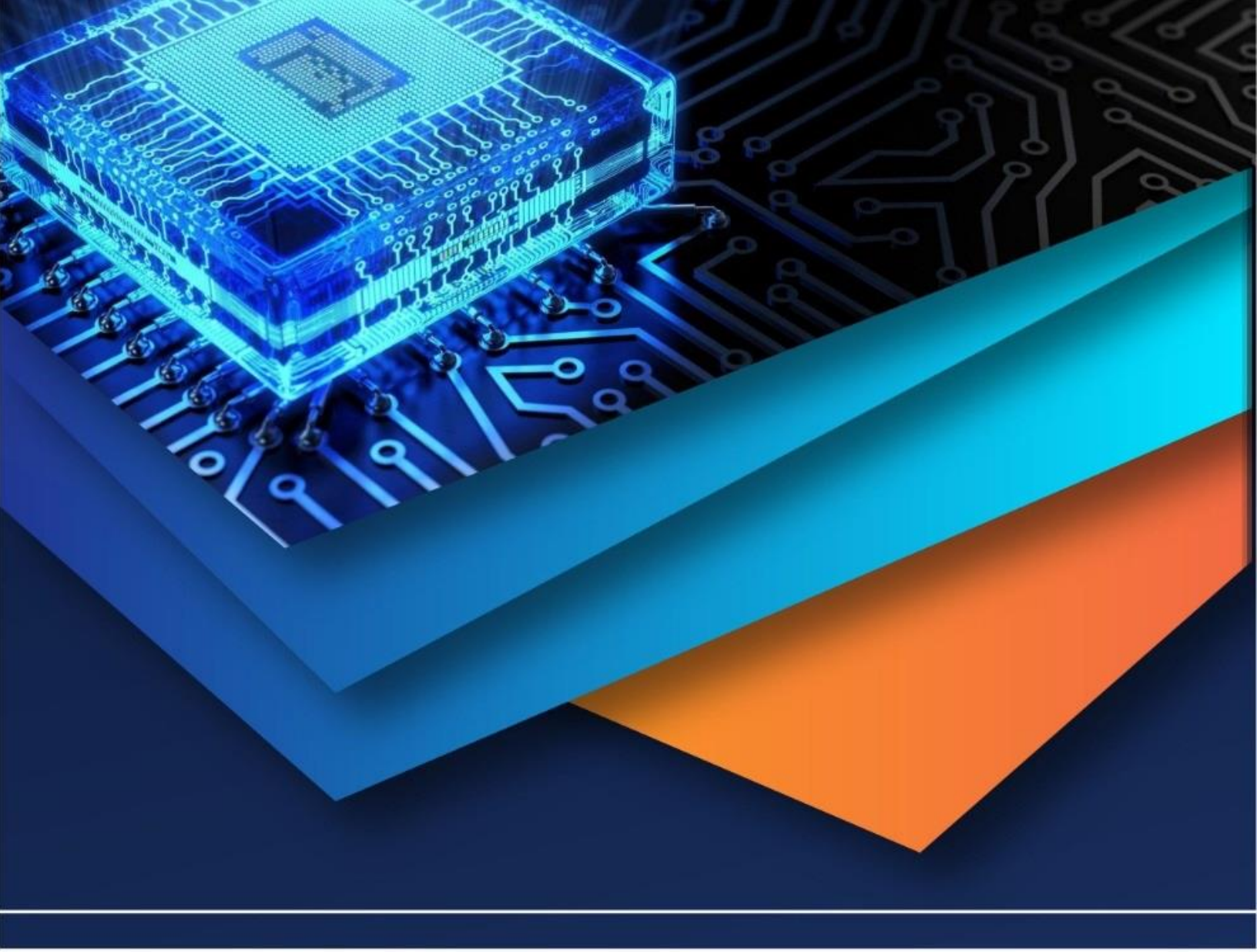
In conclusion, diamond price prediction using machine learning algorithms is an exciting field that has great potential in the diamond industry. The ability to predict the value of diamonds accurately can lead to better decision-making, reduced risks, and increased profits for buyers and sellers alike.

VI. ACKNOWLEDGMENT

We are delighted for the opportunity with great pleasure to present this paper on "Diamond Price Prediction Using Machine Learning Algorithms". We would like to seize this moment to express our gratitude to our Project Guide Prof. Amruta. A. Mankawade for providing us with exact and flawless guidance to enable us to accomplish our aspirations. We are very obliged for her support.

REFERENCES

[1] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, "Predicting stock returns by classifier ensembles," Applied Soft Computing, vol. 11, no. 2, 2011, pp 2452–2459.
[2] José M., "Implementing data mining methods to predict Diamond prices" Peña Marmolejos Graduate School of Arts and Sciences, Fordham University, Int'l Conf. Data Science ICDATA'18. https://csce.ucmss.com/cr/books/2018/LFS/CSREA2018/ ICD8070.pdf
[3] GradientBoostingRegressor with sci-kit learn [online]- https://scikitlearn.org/stable/modules/generated/sklearn.ense mble. Gradient Boosting Regressor.htm
[4] A. C. Pandey, S. Misra and M. Saxena, "Gold and Diamond Price Prediction Using Enhanced Ensemble Learning," 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, doi: 10.1109/IC3.2019.8844910, pp. 1-4.
[5] Singfat the Chu, "Pricing the Cs of diamond stones", National University of Singapore, Journal of Statistics Education Volume. https://www.tandfonline.com/doi/full/10.1080/10691898.200 1.11910 659
[6] Diamond-The most popular gemstone [online] https://geology.com/minerals/diamond.shtml
[7] Waad Alsuraihi, Ekram Al-hazmi, Kholoud Bawazeer, Hanan AlGhamdi, "Machine Learning Algorithms for Diamond Price Prediction", Publication: IVSP '20: Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, March 2020.
[8] Alexandru Niculescu-Mizil, Rich Caruana, "Predicting good probabilities with supervised Learning", Publication: learning August 2005.
[9] Supervised Machine Learning Models with sci-kit learn [online]-https://scikitlearn.org/stable/supervised_learning.html
[10] Linear, Ridge and Lasso regression with sci-kit learn [online]- https://www.pluralsight.com/guides/linear-lasso ridge-regressionscikit-learn
[11] I. ul Sami and K. N. Junejo, "Predicting future gold rates using machine learning approach."
[12] Y. Zhu and C. Zhang, "Gold price prediction based on pca ga-bp neural network," Journal of Computer and Communications, vol. 6, no. 07, p. 22, 2018.
[13] Decision tree and Random Forest regression [online]- https://towardsdatascience.com/decision-trees-and-random forests.
[14] Datasets - Diamonds dataset, Kaggle datasets repository [online] https://www.kaggle.com/shivam2503/diamonds
[15] Tovi Grossman, George Fitzmaurice, "Patina: Dynamic heatmaps for visualizing application usage", Publication: CHI April2013.https://dl.acm.org/doi/abs/10.1145/2470654.24664 42
[16] Chai T. "Root mean Square Error (RMSE) or Mean absolute error (MAE)", (NOAA Air Resources Laboratory (ARL), NOAA Center for Weather and Climate Prediction, 5830 University Research Court, College Park, MD 20740, USA; https://ui.adsabs.harvard.edu/abs/2014GMDD....7.1525C/abst ract
[17] Brownlee, J. (2018, May 22). "A Gentle Introduction to k fold Cross Validation". Online – "https://machinelearningmastery.com/k-foldcross-validation/ - Retrieved 21 October 2019.
[18] M. M. A. Khan, "Forecasting of gold prices (box jenkins approach)," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 662–670, 2013.
[19] P. K. Mahato and V. Attar, "Prediction of gold and silver stock price using ensemble models," in Advances in Engineering and Technology Research (ICAETR), 2014 International Conference on. IEEE, 2014, pp. 1–4.
[20] G. I. Webb, "Multiboosting: A technique for combining boosting and wagging," Machine learning, vol. 40, no. 2, pp. 159–196, 2000.
[21] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in Machine Learning Proceedings 1994. Elsevier, 1994, pp. 121–129.
[22] D. Banerjee, A. Ghosal, and I. Mukherjee, "Prediction of gold price movement using discretization procedure," in Computational Intelligence in Data Mining. Springer, 2019, pp. 345–356.
[23] A. C. Pandey* and D. S. Rajpoot, "Feature selection method based on grey wolf optimization and simulated annealing," Recent Patents on Computer Science, vol. XX, pp. XX–XX, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY