Analysis Comparison of K-Nearest Neighbor, Multi-Layer Perceptron, and Decision Tree Algorithms in Diamond Price Prediction

Ahya Radiatul Kamila*[1], Johanes Fernandes Andry[2] , Adi Wahyu Candra Kusuma[3], Eko Wahyu Prasetyo[4], Gerry Hudera Derhass[5]
[1,3,4]Program Studi Data Science, Universitas Bunda Mulia
[2]Program Studi Sistem Informasi, Universitas Bunda Mulia
[5]Program Studi Computer Science, Institut Pertanian Bogor
e-mail: [1]**ahyaradiatul@gmail.com**, [2]jandry@bundamulia.ac.id, [3]akusuma@bundamulia.ac.id , eprasetyo@bundamulia.com, [5]huderagerry@gmail.com

Abstract

Diamond price predictions are essential due to the high demand for these gemstones, valued as investments and jewelry. Diamonds are expensive due to their rarity and extraction process. Their prices vary depending on key factors like the diamond's inherent value and secondary factors such as marketing costs, brand names, and market trends. These variations often confuse customers, potentially leading to investment losses. This research aims to help investors determine the true price of diamonds based solely on their intrinsic value, excluding secondary factors. A machine learning approach was utilized to predict diamond prices, focusing on primary determinants. Three models such as Multi-Layer Perceptron (MLP), Decision Tree, and K-Nearest Neighbor (KNN) were compared with manual hyperparameter tuning to identify the best performing algorithm. Model performance was evaluated using Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). Among the models, KNN demonstrated the best results, achieving MAPE, MAE, and MSE values of 1.1%, 0.00038, and 〖2.687 x 10〗^(-6) respectively. This study offers valuable insights for investors by accurately predicting diamond prices based on fundamental attributes, minimizing the impact of secondary factors.

Keywords— Decision tree, Diamond Price Prediction, K-Nearest neighbor, Machine learning, Multi-Layer Perceptron

1. INTRODUCTION

Diamonds are gemstones formed from carbon elements that come from volcanoes and hundreds of meters below the earth's surface. The diamonds then melt and are brought to the earth's surface through volcanic eruptions. This stone experiences great pressure, trapped in hot temperatures. This is what makes this stone considered one of the most valued precious stones in the world since it's dense and strong [1]. This stone has a beautiful shape so it is usually used as jewelry. This stone can also be used as an investment since it's inflation-resistant. Diamonds have various colors, such as clear, white, black, purple, green pink, and blue. Of all the colors, the rarest is red, so this diamond has a more expensive price because the carbon composition is different from other diamonds. The price of diamonds depends on several factors, one of which is the 4C diamond factor, namely (carat, cut, color, and clarity). Apart from that, several supporting factors determine the price of diamonds, including brand name, marketing costs, designer name, location of diamond shop, etc. This causes differences in diamond prices. On the other hand, most customers do not have basic knowledge about the product, so they cannot know for sure whether the price offered is appropriate or not. However, in investing, investors are required to know the real price of the investment instrument they choose, so they can predict the risk and return of investment well [2]. To solve this problem, a system is needed that can predict the original price

of a diamond based on the main factors that determine the price of a diamond (carat, cut, color, and clarity) without taking into account the supporting factors of the diamond price.

Several studies have been carried out to predict diamond prices using several different algorithms. Sharma et al [3] proposed research that uses the Multiple Linear Regressor (MLR) algorithms to determine the correlation between the determining factors of diamond prices. The result obtained from this research is that the weight of a diamond does not have a linear relationship with its price. Heavier diamonds are not necessarily more expensive than lighter diamonds. This is because several other factors influence the price of diamonds. The research proposed by [4] compares model performance using three machine learning algorithms, Linear Regression, Decision Tree, and K-nearest neighbor. From this research, it was found that the best model performance was produced using the Decision Tree algorithm with an accuracy of 88% or an average percentage of absolute error of around 12%. A study conducted by [5] elucidated the importance of certain factors in determining diamond prices. Carat, which is a unit for measuring the weight of a diamond, was found to be the most significant factor, followed by width, clarity, and color. The research conducted by comparing 5 machine learning algorithms, among Linear regression, Gradient descent, Random Forest regression, Polynomial regression, and Neural network with Random Forest regression turned out as the best performance.

This research aims to identify the best algorithm for predicting diamond prices by comparing several machine learning algorithms. By identifying the most effective algorithm, this study aims to generate the highest accuracy model for predicting diamond prices. The results of this best-performing algorithm can be used for business purposes, such as optimizing pricing strategies, reducing financial risks, and improving inventory management efficiency in the diamond industry. With more accurate price predictions, businesses can make more precise pricing decisions, enhance competitiveness in the market, and provide added value to customers investing in diamonds. In this research, the author compares three machine learning algorithms, Multi-Layer Perceptron (MLP), Decision Tree, and K-Nearest Neighbor (KNN). These algorithms were selected for their diverse approaches to modeling data and their relevance in regression tasks. MLP, as a type of neural network, is capable of capturing complex, non-linear relationships between features, making it suitable for datasets where interactions among variables are intricate. The decision tree is a rule-based model that excels in interpretability and can effectively handle both categorical and numerical data, making it a versatile choice for regression problems. KNN, on the other hand, is a distance-based algorithm that predicts values by averaging the outcomes of the nearest neighbors, which is particularly useful for understanding localized patterns in the data. Of these three machine learning algorithms, an algorithm with a regression approach is used, thus MLP Regressor, Decision Tree Regressor, and KNN Regressor are used. Regression is a technique for determining the relationship between dependent variables and independent variables so that they can be predicted [6]. This is done using the same exploratory data analysis and data pre-processing steps until we get results comparing model performance that truly illustrates the ability of each machine learning algorithm to predict diamond prices. Exploratory data analysis is a stage of data analysis that aims to see the meaning of the dataset and potential problems in it. This stage is usually carried out by looking at data visualization [7]. Data visualization can represent data or information using graphs, charts, or other visual formats [8]. So, by using data visualization to analyze datasets, meaning, patterns/trends, data anomalies, etc. can be identified. From the best machine learning algorithm, diamond price predictions will be produced which can be used as a reference for customers when investing in diamonds. In other words, this research contributes to the business and investment sector in determining the price of investment instruments following the actual price without being influenced by supporting factors that cannot be calculated/predicted.

2. RESEARCH METHODS

The research methodology in this paper is divided into three parts. The first section looks at the phases of data processing and the dataset that was used. The second part examines the

correlation between the features used. The third section examines the data pre-processing stages used to improve model performance. This research focuses on the results of comparing the performance of three machine learning models, Multi-Layer Perceptron, Decision Tree, and K-Nearest Neighbor. Apart from that, this research also uses several methods in the data pre-processing stage which are used to maximize the performance results of the machine learning model. Figure 1 is a block diagram of this research.

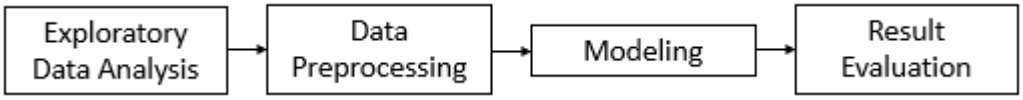

Figure 1. Block Diagram

In this research, the first step taken was the data imputation stage. This stage is carried out by reading the dataset in .csv format using the Python programming language. After that, the exploratory data analysis stage was carried out. This stage is the stage of raw data analysis which is carried out to understand the dataset so that the author understands the research dataset and determines decisions regarding what should be done at the next stage [9]. The next stage is the data pre-processing. This stage is one of the most crucial stages in this research. At this stage, several methods are carried out, such as changing raw data into data in a more understandable format, ensuring that the dataset used is clean (does not contain irrelevant data), etc. [10]. This is used to improve model performance, reduce computational costs, and facilitate interpretation of the dataset [11]. After the data has gone through the pre-processing stage and is deemed suitable for processing, the data is then processed using a machine learning model so that it can produce prediction results. The prediction results from machine learning algorithms are then evaluated using model performance metrics. The model performance metrics used in this research are the Average Percentage of Absolute Error, Average Absolute Error, and Average Squared Error. Then the model performance is compared to obtain an algorithm that produces the best model performance with the smallest error value.

2.1. Dataset

This research dataset is sourced from Kaggle, which is an open-source data with 53,940 rows and 10 columns (features) with the dataset title 'diamonds.csv'. The features used in the diamond's dataset consist of nine input columns (Independent Features) and one output column (Dependent Feature) with the column title 'price'. This dataset is a structured dataset, since it has an organized format, with data stored in a table consisting of rows and columns, where each column represents a specific feature, and each row represents an entity or data point. The the context of the "diamonds.csv" dataset, the column consists of attributes such as carat, cut, color, clarity, depth, table, price, x, y, and z, while the rows represent individual diamonds with values corresponding to these attributes.

	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
...	...	...	...	...	...	...	...	...	...	...	...
53935	53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 11 columns

Figure 2. Dataset

From Figure 2. The carat feature, a numerical variable, represents the weight of the diamond in carats and is one of the most significant factors influencing its price. The categorical cut feature indicates the quality of the diamond's cut and is categorized into five levels: Fair, Good, Very Good, Premium, and Ideal. Another categorical feature is color, which describes the diamond's color grade. Similarly, the clarity feature categorizes the diamond based on the presence of inclusions or blemishes, with levels from I1 (most inclusions) to IF (Internally Flawless, the highest clarity). The dataset also includes numerical features such as depth, which is the depth percentage of the diamond calculated as the ratio of its height to its average diameter, and table, which represents the width of the diamond's table as a percentage of its diameter. The physical dimensions of the diamond are captured in the numerical features x, y, and z, representing the length, width, and height of the diamond in millimeters, respectively. The dependent feature, price, is a numerical variable representing the cost of the diamond in US dollars and serves as the target variable for prediction in this study. These features collectively provide a comprehensive basis for understanding and modeling the factors influencing diamond prices.

2.2. Exploratory Data Analysis

Exploratory Data Analysis is the first step needed in research since it can help the writer recognize potential problems from the data, avoid bias, and plan the next steps that will be taken in the pre-processing stage.

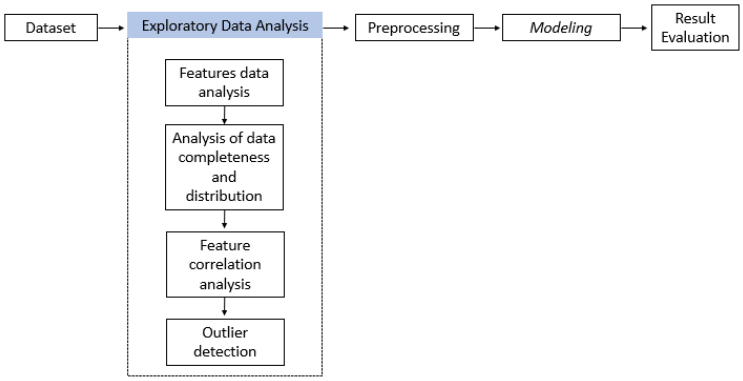

Figure 3. Block Diagram of Exploratory Data Analysis Stage

Figure 3 represents the sequential process undertaken in this research to analyze and model diamond price predictions. The workflow begins with the *Dataset* phase, which involves sourcing the data and understanding its structure. Following this is the *Exploratory Data Analysis (EDA)* phase, which is broken into several key steps:

1. Features Data Analysis - This step involves examining the dataset's attributes to understand their nature, distribution, and potential significance in the model.
2. Analysis of Data Completeness and Distribution - In this stage, the dataset is checked for missing values, and the distribution of features is analyzed to ensure data integrity and suitability for modeling.
3. Feature Correlation Analysis - This step assesses the relationships between features to identify which attributes are strongly correlated with the target variable (*price*) and with each other.
4. Outlier Detection - Outliers in the data are identified and handled appropriately, as they can significantly affect the performance of machine learning models.

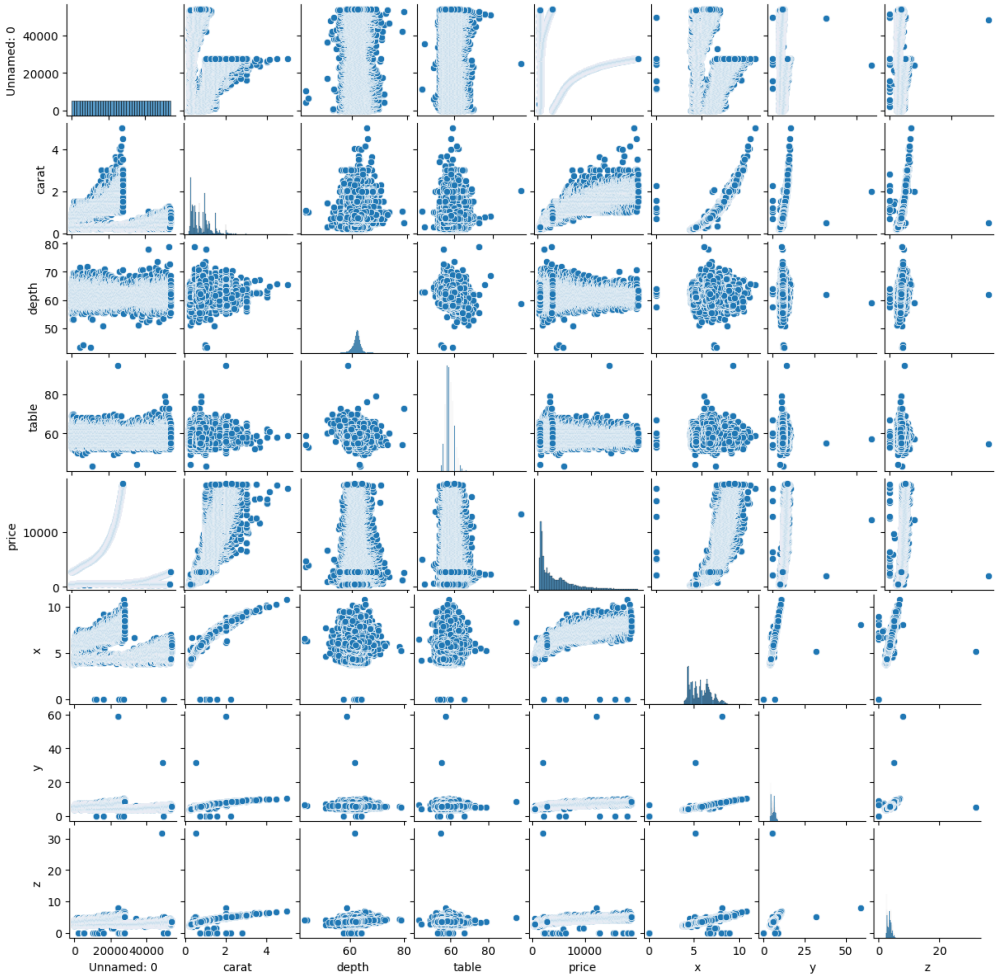

Figure 4. Pairplot

Figure 4 represents the pair plot that was utilized to visualize the relationships between numerical features in the dataset. A pair plot is a grid of scatterplots and histograms that provides insights into how each pair of features correlates, as well as the distribution of individual features. By plotting these relationships, the pair plot helps identify patterns, trends, or clusters within the data. Additionally, it serves as an essential tool for observing data distribution and detecting

potential anomalies, such as outliers or irregular patterns, which may affect the model's performance if not addressed.

On the other hand, to represent the correlations between the dataset's features, we use a heatmap (*Figure 5*). A heatmap is a powerful visualization tool that displays the correlation matrix in a grid format, where the color intensity indicates the strength of the correlation between different features (independent with independent features or independent with dependent features). The correlation between independent features and dependent features indicates the impact of a feature on its output. Meanwhile, the relationship between independent features and other independent features can indicate data redundancy. In this study, visualization using a heatmap with a triangle correlation heatmap type is employed to depict the correlations between features.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\hat{x})(y_i-\hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\hat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\hat{y})^2}} \quad (1)$$

Where :
x_i = Independent feature
y_i = Dependent feature
r_{xy} = Correlation between features

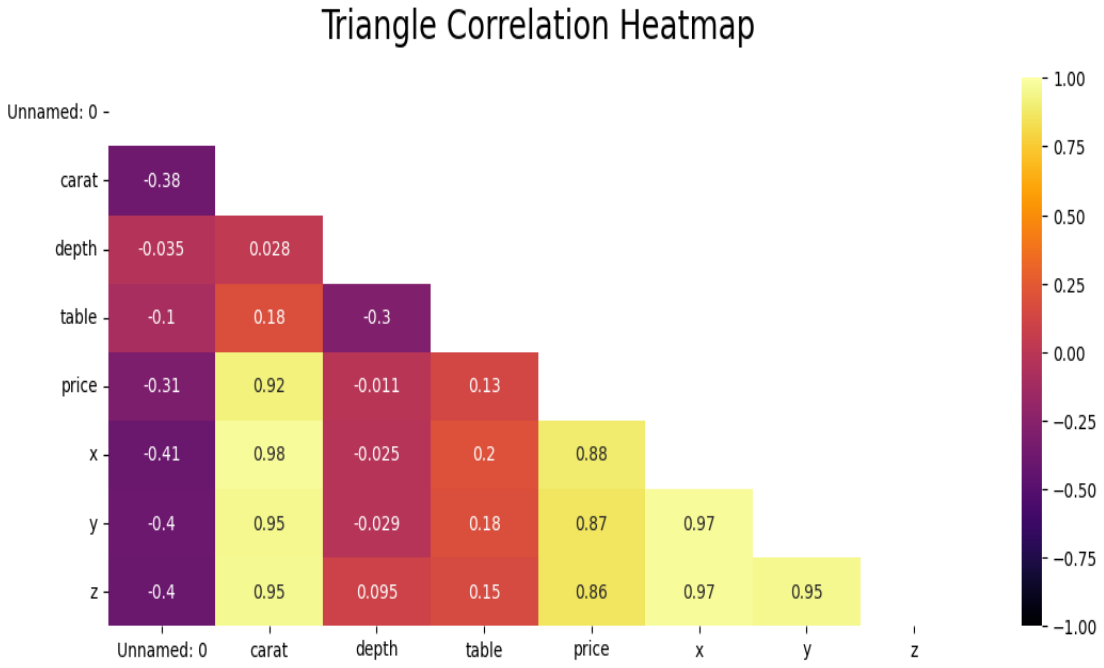

Figure. 5 Heatmap Correlation

The calculated value of r_{xy} ranges from -1 to 1. If the correlation result approaches 1, it indicates a positive correlation between the two features, meaning they are directly proportional. If the correlation result approaches -1, it suggests a negative correlation, meaning the two features are inversely proportional. Conversely, if the correlation result is close to 0, it implies that the two features do not correlate with each other, meaning that the value in feature A does not affect the value of the feature.

2.3. Preprocessing

The data preprocessing method is a stage that is performed before data can be processed by a machine learning model to generate prediction outputs. The objective of this step is to prepare

raw data into a more suitable form for analysis and modeling [12]. This adjustment is made to make the dataset easier to interpret and to produce a better model performance. This stage involves remodeling and transforming raw data into a more efficient format [13]. In this study, the author proposes four data preprocessing stages, which can be seen in Figure 6.

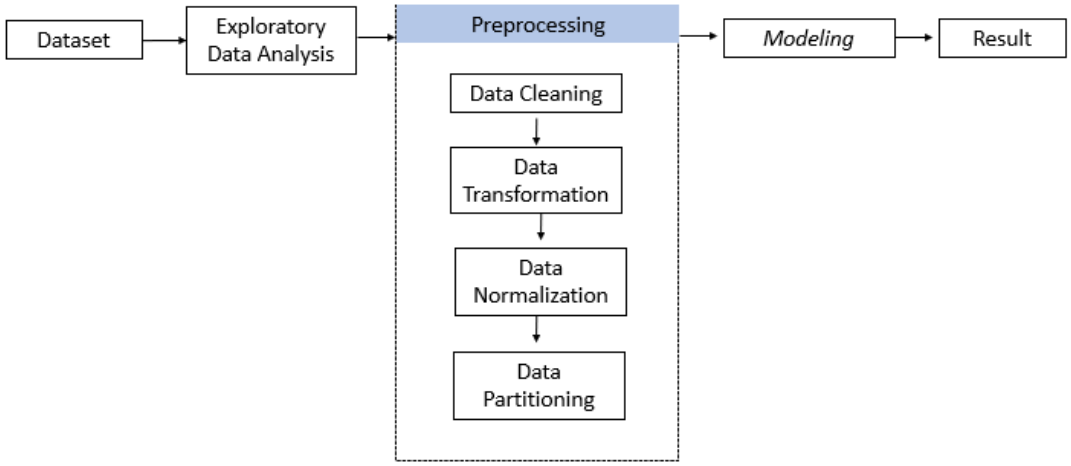

Figure 6. Block Diagram of Preprocessing Stage

Data cleaning is the process used to identify inaccurate, incomplete, or incorrect data, which helps improve the accuracy of the model. In this stage, irrelevant or inconsistent data that does not align with the pattern and characteristics of the dataset is discarded, a process commonly referred to as outlier removal. This step enhances the quality of analysis and the performance of machine learning models. Furthermore, data transformation involves converting features in the dataset into a more suitable form, particularly when the dataset contains categorical features. Since most machine learning algorithms cannot work with categorical data types, data transformation is necessary to convert them into numeric types. In this study, the OneHotEncoder method is used for this purpose. Additionally, data normalization aims to change the range of data values, which helps speed up the learning process in machine learning and facilitates data analysis. Several methods exist for normalizing data, and in this study, the MinMaxScaler method is applied. Finally, data partitioning divides the dataset into training and test data. The training data is used to train the model, while the test data is used to evaluate the model's performance. This method allows for assessing the model's ability to generalize, and in this study, the data is split into 70% training data and 30% test data.

2.4. Modeling

Three different machine learning algorithms with different working methods are compared in this research. This is done using the exactly same process of exploratory data analysis and preprocessing stages, thus the results in model performance reflect the ability of each algorithm to solve the diamond price prediction problem. After obtaining the model performance of each machine learning algorithm, a comparison is made between the three algorithms. This is done to determine the algorithm with the best model performance. The hyperparameters of the model used in this study are determined by trial and error. In this study, the author tunes the hyperparameters of each algorithm to improve model performance, since choosing the wrong hyperparameters can result in an inappropriate model. Specifically, the tuning process involves testing various parameter ranges for each algorithm. For the Multi-layer Perceptron (MLP), parameters such as the number of hidden layers, the number of units per hidden layer, and the learning rate were tuned. For the Decision Tree algorithm, hyperparameters like the maximum depth, minimum samples split, and minimum samples leaf were adjusted. Lastly, for K-Nearest

Neighbors (KNN), the number of neighbors and distance metrics were tested. These hyperparameter adjustments were made to optimize the models for better accuracy and performance.

2.4.1. Multi-Layer Perceptron (MLP)

As an enhancement of the previous algorithm, Single Layer Perceptron, which is the precursor to Neural Network algorithms, the Multi-Layer Perceptron algorithm constructs hyperplanes to separate different data sets in high-dimensional space. Thus, this algorithm can handle the problem of separating a set of linear data well [14]. However, this algorithm has weaknesses in generalizing non-linear datasets. Therefore, Multi-Layer Perceptron (MLP) was created.

2.4.2. Decision Tree

The decision tree is a supervised learning algorithm that can be used in both regression and classification problems. In solving regression problems, a Decision Tree Regressor is used, while a Decision Tree Classifier is used in classification. This algorithm has a structure that resembles a decision tree, consisting of nodes that form a rooted tree and have leaves. Decision leaves are the final stage that will determine the prediction result [15].

2.4.3. K-Nearest Neighbours

K-Nearest Neighbor commonly known as KNN belongs to the group of instance-based learning. This algorithm works by searching for several objects that are closest to other objects [16]. The idea of this algorithm is to assign new data to the class group of data whose majority neighbors are K nearest. The first step when using this algorithm is to determine the number of neighbors (K). Next, the KNN algorithm will calculate the distance from the test data feature to all the features of the training data that have been obtained. After that, the K features of the training data with the closest distance to the test data feature are selected, and then the prediction category of the test data is made based on the specified K value [17].

3. RESULTS AND DISCUSSIONS

Machine learning algorithms in forecasting may not always yield accurate results. Therefore, evaluation is required, which can be done by comparing the forecasted results with actual occurrences to determine the adequacy of the model performance [18]. In this study, three metrics performance models were used to measure the model's performance, Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Mean absolute percentage error is the metric that represents the average value of absolute differences between predicted and actual values. These metrics can assess the accuracy of the forecasted values compared to the actual values which are expressed as a percentage [19]. This metric assesses the accuracy of the forecasted values compared to the actual values, as shown in the following equation:

$$\text{MAPE} \quad : \frac{1}{n}\sum \left|\frac{y-\hat{y}}{\hat{y}}\right| x\ 100\% \qquad (2)$$

Mean absolute error is another metric used to evaluate model performance in this study. The calculation result of these metrics indicates the average absolute error between actual and forecasted values. Based on equation 3, MAE calculates the average error by assigning equal weights to all data points, which leads to being more intuitive in providing the average error of the entire dataset [20]

$$\text{MAE} \quad : \frac{1}{n}\sum |y-\hat{y}| \qquad (3)$$

Mean squared error is a metric used to measure the average of the squares of the errors or differences between predicted and actual [21] values in a regression problem. It provides a way to quantify the amount of variation or dispersion in the predictions made by the model, with a lower MSE indicating that the model's predictions are closer to the actual values. MSE is

calculated by taking the average of squared differences between the predicted and actual values for each data point. It is a popular metric in evaluating model performance by providing a numerical value that represents the quality of the model's predictions [22]

MSE : $\frac{1}{n}\sum(y-\hat{y})^2$ *(4)*

Where :

y : actual value

$\hat{y}$: predicted value

In the results section, these metrics (MAPE, MAE, and MSE) are essential for evaluating model usability. A lower MAPE indicates better accuracy of the model's predictions in percentage terms with an intuitive interpretation [23], while a lower MAE reflects fewer absolute errors between the actual and predicted values. MSE, being sensitive to larger errors due to its squaring of differences, helps identify models with significant prediction errors. These metrics collectively provide a comprehensive evaluation of the model's performance, guiding decisions on model selection and optimization for practical use.

3.1. Multi-Layer Perceptron (MLP)

In this study, the Multi-Layer Perceptron Regressor algorithm was used as one of the algorithms, and its performance was compared with other algorithms in the case of diamond price prediction. Since the output of this study is a continuous number, this study uses a regression approach with a regressor algorithm to predict diamond prices. The performance of the model with the MLP Regressor algorithm in this study can be seen in Table 1.

Table 1. Loss and Accuracy Values using the Optimizer at Epoch 1-100

Metrics Performance Model	**Result**
MAPE	40%
MAE	0.03356
MSE	0.00320

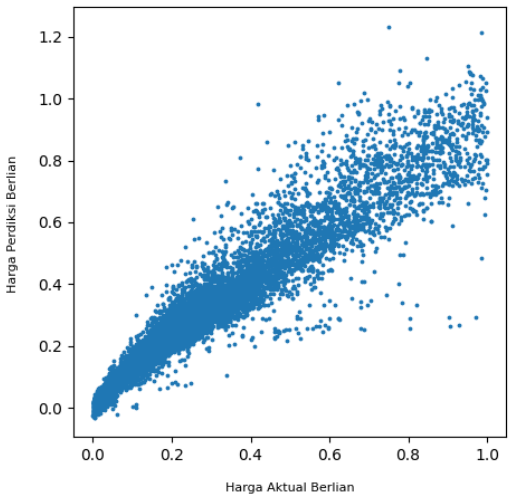

Figure 7. Comparison of Predicted and Actual Prices of Diamonds with MLP Regressor

The figure titled "Comparison of Predicted and Actual Prices of Diamonds with MLP Regressor" illustrates the performance of the Multi-layer Perceptron (MLP) Regressor model in predicting diamond prices. This figure will be compared with the results from other models, specifically the Decision Tree and K-Nearest Neighbors (KNN) models. The MLP Regressor in this plot is being compared against these models to assess which model performs better in predicting diamond prices. In particular, the data shown in this plot represents the case where the

values are most dispersed, indicating that the model's predictions have the greatest deviation from the actual values. This highlights the regions where the MLP model has the most significant prediction errors compared to the actual diamond prices. The comparison between these models helps determine which one produces more accurate and reliable predictions for the given dataset.

3.2. Decision Tree

The Decision Tree algorithm used in this study is the Decision Tree Regressor. This algorithm serves as the second model whose performance is compared with the other two algorithms, namely the Multi-layer Perceptron (MLP) and K-Nearest Neighbors (KNN). The performance of the Decision Tree model is presented in Table 2. In terms of model accuracy, the Decision Tree algorithm shows lower error rates compared to the MLP Regressor, but its error rate is higher than that of the KNN model. This suggests that while the Decision Tree performs relatively well, it does not achieve the level of accuracy seen in the KNN model.

Table 2. Performance Model with Decision Tree Regressor

Metrics Performance Model	Result
MAPE	15%
MAE	0.01945
MSE	0.00156

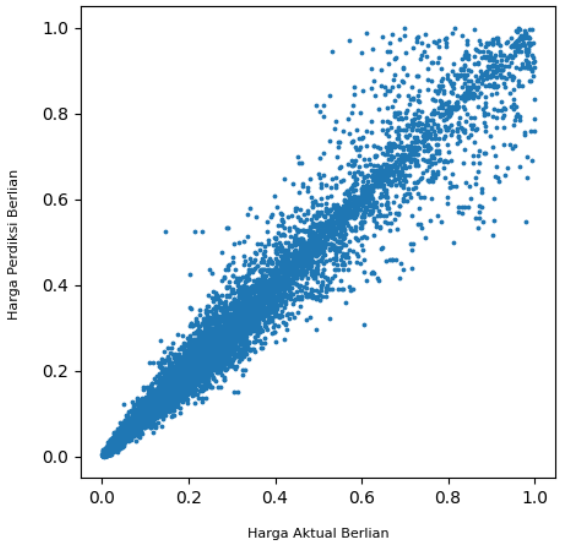

Figure 8. Comparison of Predicted and Actual Prices of Diamonds with Decision Tree Regressor

This comparison is further illustrated in Figure 8, which shows a plot representing data with less dispersion than the MLP Regressor plot. The more compact distribution of data in the Decision Tree plot indicates that the model's predictions are closer to the actual values than in the MLP case, although they still exhibit larger deviations compared to the KNN model. This visualization highlights the Decision Tree's relative performance in the context of the three algorithms being compared.

3.3. K-Nearest Neighbor

The K-nearest neighbor algorithm is the last algorithm the author uses to predict diamond prices. Similar to the two previous algorithms, the K-nearest neighbor algorithm used in this study is the K-nearest neighbor Regressor. The number of neighbors (K) used in this study is 5 nearest neighbors. The model performance generated using the K-Nearest Neighbor Regressor algorithm is listed in Table 3.

Among the three algorithms, the KNN model demonstrates the lowest error rates, outperforming both the MLP Regressor and the Decision Tree Regressor in terms of prediction accuracy. This suggests that the KNN algorithm provides more reliable and precise predictions for diamond prices in this dataset. The relatively smaller error margin highlights the strength of the KNN model in capturing the relationships between the features and the target variable. This finding is consistent with the results presented in the study, where KNN provides the most accurate predictions, making it the most suitable model for this particular task.

Table 3 Performance Model with K-Nearest Neighbor Regressor

Metrics Performance Model	**Result**
MAPE	1.1%
MAE	0.00038
MSE	2.68×10^{-6}

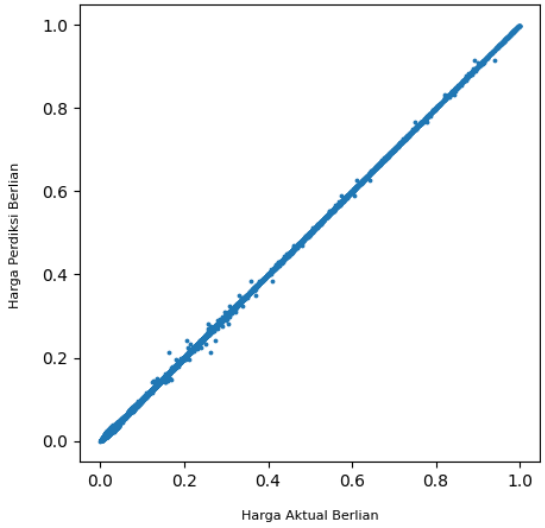

Figure 9. Comparison of Predicted and Actual Prices of Diamonds with K-Nearest Neighbor Regressor

Figure 9 provides a visualization of the error generated by the K-Nearest Neighbor (KNN) Regressor, showcasing the model's performance with the most accurate predictions. In this figure, the data distribution is the least dispersed compared to the MLP and Decision Tree models, reflecting the KNN model's ability to generate predictions that are very close to the actual values. The smaller spread of data points in this plot indicates that the KNN model has minimal prediction errors, highlighting its superior accuracy. This visualization underscores the KNN model's effectiveness in providing precise estimates of diamond prices, with the error being the smallest among the three algorithms compared. The compactness of the data points reinforces the conclusion that the KNN Regressor is the most reliable model for predicting diamond prices in this study.

4. CONCLUSION

Based on the experiments conducted, the author can conclude that the selection of a suitable machine learning algorithm for the dataset used significantly affects the model's performance, as evidenced by the performance results of the machine learning models. Among the three algorithms tested (K-Nearest Neighbor, Multi-Layer Perceptron, and Decision Tree), K-Nearest Neighbor demonstrated the smallest error value, indicating that it produced the best model performance in predicting diamond prices with the dataset used in this study. This highlights the

importance of choosing the right algorithm to achieve the most accurate predictions for a given problem.

The contributions of this study include providing a detailed comparison of multiple machine learning algorithms, evaluating their performance in predicting diamond prices and highlighting the significance of model selection. The findings contribute valuable insights into how different algorithms behave when applied to the same dataset, offering a benchmark for future research in this area. In the business context, this research provides valuable guidance for companies in the diamond industry by showcasing the potential of machine learning in improving pricing strategies and forecasting models. By identifying the most effective algorithm, such as K-Nearest Neighbour, businesses can optimize pricing accuracy, reduce financial risks, and improve inventory management. This study encourages the adoption of advanced analytics, facilitating data-driven decision-making that enhances competitive advantage in the market. Additionally, the research emphasizes the importance of data preprocessing and hyperparameter optimization, which can be tailored to specific business needs, further improving the accuracy and efficiency of the models. Ultimately, the study contributes to the digital transformation of businesses by enabling more precise predictions, better alignment with market trends, and more effective pricing strategies.

For future studies, the author suggests further research development by utilizing hyperparameter optimization algorithms for each model to achieve more optimal performance results. Hyperparameter tuning could further enhance the predictive power of the models. Additionally, comparing the data preprocessing methods of each machine learning algorithm used would help determine the most suitable data preprocessing method for the dataset, enhancing the accuracy and efficiency of the models. Such improvements could lead to more robust and reliable predictions in similar real-world applications, ultimately driving better decision-making processes in industries where diamond price prediction is crucial.

REFERENCES

[1] Sonia, "Diamond Price Prediction Using Machine Learning Algorithms", *International Journal of Multidisciplinary Educational Research,* vol. 12, pp.99-106, June 2023.

[2] W. Alsuraihi, E. Al-Hazmi, K. Bawazeer, and H. Alghamdi, "Machine Learning Algorithms for Diamond Price Prediction," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Mar. 2020, pp. 150–154. doi: 10.1145/3388818.3393715.

[3] G. Sharma, V. Tripathi, M. Mahajan, and A. K. Srivastava, "Comparative analysis of supervised models for diamond price prediction," in *Proceedings of the Confluence 2021: 11th International Conference on Cloud Computing, Data Science and Engineering*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 1019–1022. doi: 10.1109/Confluence51648.2021.9377183.

[4] A. A. Mankawade, C. Kokate, K. Soman, A. Mohite, A. Vispute, and O. More, "Diamond Price Prediction Using Machine Learning Algorithms," *Int J Res Appl Sci Eng Technol*, vol. 11, no. 5, pp. 4867–4871, May 2023, doi: 10.22214/ijraset.2023.52741.

[5] H. Zhang, "Prediction and Feature Importance Analysis for Diamond Price Based on Machine Learning Models," *Advances in Economics, Management and Political Sciences*, vol. 46, no. 1, pp. 254–259, Dec. 2023, doi: 10.54254/2754-1169/46/20230347.

[6] J. F. Andry, F. M. Silaen, H. Tannady, and K. H. Saputra, "Electronic health record to predict a heart attack used data mining with Naïve Bayes method," *International Journal*

of Informatics and Communication Technology (IJ-ICT), vol. 10, no. 3, p. 182, Dec. 2021, doi: 10.11591/ijict.v10i3.pp182-187.

[7] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," Sep. 01, 2021, *Springer*. doi: 10.1007/s42979-021-00765-8.

[8] J. Fernandes Andry, H. Tannady, I. Ivana Limawal, G. Dwinoor Rembulan, and R. Farady Marta, "Big Data Analysis on Youtube with Tableau," *J Theor Appl Inf Technol*, vol. 99, p. 22, 2021, [Online]. Available: www.jatit.org

[9] M. B. Courtney, "Exploratory Data Analysis in Schools: A Logic Model to Guide Implementation," *International Journal of Education Policy and Leadership*, vol. 17, no. 4, May 2021, doi: 10.22230/ijepl.2021v17n4a1041.

[10] A. Radiatul Kamila and A. Subiantoro, "Coronary Heart Disease Detection Using a Combination of Adaptive Synthetic Sampling Approach and Stacking Method on Imbalanced and Incomplete Dataset", *International Engineering Student Conference*, June 2022.

[11] J. M. Waworundeng *et al.*, "Sentiment Analysis of Online Lectures Tweets using Naïve Bayes Classifier Analisis Sentimen Tweet Kuliah Online menggunakan Naïve Bayes Classifier," *Cogito Smart Journal* |, vol. 8, no. 2, p. 2022.

[12] F. F. Tampinongkol, R. Ilham, A. R. Kamila, Y. Purnomo, C. Herdian, S.Virginia, "Deteksi Ciri Link Phishing Menggunakan Algoritma Random Forest Untuk Meningkatkan Keamanan Cyber", *Techno Xplore Jurnal Ilmu Komputer dan Teknologi Informasi*", vol. 9, no.2, 2024.

[13] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," Mar. 29, 2021, *Frontiers Media S.A.* doi: 10.3389/fenrg.2021.652801.

[14] Y. Qin, C. Li, X. Shi, and W. Wang, "MLP-Based Regression Prediction Model For Compound Bioactivity," *Front Bioeng Biotechnol*, vol. 10, Jul. 2022, doi: 10.3389/fbioe.2022.946329.

[15] A. Saleh, and M. Maryam, "Pemanfaatan Teknik Data Mining Dalam Menentukan Standar Mutu Jagung The Utilization Data Mining Technique in Determining the Quality Standard of Corn," *Cogito Smart Journal* |, vol. 5, no. 2, p. 171. 2019

[16] E. Hasmin, C. Susanto, K. Aryasa, U. Dipa Makassar, and J. Perintis Kemerdekaan Km, "Sistem Pakar Prediksi Penyakit Diabetes Menggunakan Metode K-NN Berbasis Android Expert System for Predicting Diabetes Using the Android-Based K-NN Method," *Cogito Smart Journal* |, vol. 8, no. 2. 2022.

[17] A. Pamuji, "Performance of the K-Nearest Neighbors Method on Analysis of Social Media Sentiment," *JUISI*, vol. 07, no. 01, 2021.

[18] A. Sujiana and U. Budiyanto, "Prediksi jumlah Produksi Perakitan Komponen Menggunakan ANFIS Yang Dioptimasi Dengan Algoritma K-Means Prediction of Component Assembly Production Using ANFIS Optimized With K-Means Algorithm," *Cogito Smart Journal* |, vol. 9, no. 2, 2023.

[19] I. Nabillah and I. Ranggadara, "Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut," *JOINS (Journal of Information System)*, vol. 5, no. 2, pp. 250–255, Nov. 2020, doi: 10.33633/joins.v5i2.3900.

[20] A. A. Suryanto, A. Muqtadir, and S. Artikel, "Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi" *Jurnal Sains dan Teknologi*, no. 1, p. 11, 2019.

[21] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean Squared Error, Deconstructed," *J Adv Model Earth Syst*, vol. 13, no. 12, Dec. 2021, doi: 10.1029/2021MS002681.

[22] H. Mustafidah and S. N. Rohman, "Mean Square Error pada Metode Random dan Nguyen Widrow dalam Jaringan Syaraf Tiruan Mean Square Error on Random and Nguyen Widrow Method on Artificial Neural Networks", 2023, doi: 10.30595/sainteks.v20i2.19516.

[23] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.