Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction

Shafilah Ahmad Fitriani
Department of Informatics
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
shafilah.f@students.amikom.ac.id

Yuli Astuti
Department of Informatics
Management
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
yuli@amikom.ac.id

Irma Rofni wulandari
Department of Information System
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
irma@amikom.ac.id

2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE) | 978-1-6654-0544-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISMODE53584.2022.9742936

Abstract— Diamonds are the most expensive, rarest, and most complex gemstones globally. Diamond investing is a new lifestyle; however, diamond prices fluctuate and are difficult to predict. Predicting the price of diamonds can be done using a regression technique because the price is continuous. Regression is part of the field of machine learning. This study aims to find the most efficient and accurate model. The models used to predict diamond prices are k-Nearest Neighbors (k-NN) and Least Absolute Shrinkage and Selection Operator (LASSO). The process is carried out by selecting features, considering the value of k from k-NN and alpha from LASSO to ensure optimal accuracy. The data of this research is public and taken from Kaggle. The number of datasets is around 54000 data and is divided into training data by 80% and testing data by 20%. The results showed that k-NN had the highest accuracy of 0.9066 compared to LASSO, which was 0.8801. Meanwhile, the RMSE level shows that k-NN has the smallest value, 926.06, compared to LASSO, 1049.59.

Keywords—Machine learning,* Diamonds*, k-NN, LASSO

I. INTRODUCTION

Diamond is one of the rarest and hardest naturally occurring minerals of carbon [1]. Diamonds have been used as gemstones for centuries because when adjusted for aspect, diamonds have a characteristic 'fire' due to their high refractive index [2]. In society, diamonds are trending as jewelry as gifts at particular moments. Not only as a gift but diamonds are also seen as an alternative investment other than gold because their value can provide significant profits and increases. As a customer, there are always doubts about the right time to buy or sell diamonds to invest in.

In the diamond trading sector, buyers or investors face difficulties in predicting the price of diamonds due to differences in the shape, size, and purity of diamonds [3]. Many models and applications have been implemented to predict the future price of this diamond using machine learning. Machine learning is divided into two categories, namely supervised and unsupervised. Supervised learning algorithms use the general principle of practical examples for prediction [4] or forecasting. Several machine learning models are used to predict future diamond prices, such as linear regression, random forest regression, polynomial regression, gradient descent, AdaBoost Regressor, ElasticNet, Gradient-Boosting Regressor, LASSO, and neural networks. Previous studies have compared methods that produce the best accuracy and results on the random forest regression algorithm model [1][3]. Therefore, this study aims to find the most efficient and highest accuracy model for predicting diamond prices from two machine learning algorithms, namely k-NN and LASSO, using the feature selection provided by scikit-learn.

Before performing feature selection and the algorithm model to be used, an important step must be done, namely preprocessing data, which has an essential effect on the machine learning supervised algorithm [5]. These crucial steps are:

1. Data Cleaning

The dataset has missing value or noise data. This is because the data collection process is not perfect, resulting in irrelevant or missing data. Inconsistencies and incompleteness caused by human or system errors can affect machine learning to be built [6]. So that data cleaning needs to be done in order to find duplicates in the dataset [7].

2. Categorial Encoding

Most of the datasets are categorical or strings. In some cases, building a machine learning model requires numerical data converted to categorical data into numeric data one-hot encoding or Label-Encoding is performed.

In one-hot encoding, each category is mapped to binary 0 or 1, while Label-Encoding, each label is converted to a numeric value or integer [8]. Doing categorical encoding is very easy via the scikit-learn or pandas libraries.

3. Scale features

Scale features is a pre-processing method; to implement the k-NN and LASSO algorithms, it is necessary to use data that has many features. Several methods exist in the Scale features, such as a standard, min-max, and robust scaler.

II. RESEARCH METHODOLOGY

The stages of the research method carried out in this study include data collection, data cleaning, pre-processing, feature selection, and modeling.

A. Data Collection

The machine learning model in this study uses the Diamonds dataset taken from the Kaggle website. The dataset contains prices and other attributes of nearly 54,000 diamonds. The dataset is in the form of CSV files. Features used with the following description:

1. Price

price in US dollars (\$326—\$18.823)

2. Carat

   carat weight of the diamond (0.2—5.01)

3. Cut

   cut quality (Fair, Good, Very Good, Premium, Ideal)

4. Color

   diamond color, from J (worst) to D (best)

5. Clarity

   a measure of how clear a diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))

6. x: length in mm (0—10.74)

7. y: width in mm (0—58.9)

8. z: depth in mm (0-31.8)

9. depth

   total depth percentage = z / mean(x,y) = 2*z / (x+y)(43—79)

10. table

    the width of the top of the diamond relative to the widest point (49-95) price : US dollars (\$326—\$18,823)

B. Data Cleaning

In this process, data cleaning is carried out. This process includes checking the data, whether the data is empty, inconsistent, or duplicated. There is no missing data in this dataset. Some of the data is categorical data, so changes must be made to numerical data using categorical encoding techniques.

C. Pre-processing

Pre-processing of data often significantly impacts supervised machine learning algorithms [9]. This study does not support null values and categorical data. There are no null values or missing values in this diamond dataset, but three columns are included in the category data, namely cut, color, clarity. So to handle this case, Label-Encoding is done to convert numeric values using the libraries available in scikit-learn. After performing categorical encoding, the dataset will be divided into training and test data. Training data is used to train and adjust the model, while testing data is to test the trained model. The model is evaluated for accuracy using testing data [10].

In this Diamonds dataset, the author divides the dataset into two parts, a training dataset of 80% and a testing dataset of 20% using the scikit-learn library.

The next step is to get the best features from the dataset. The author uses the scikit-learn library, namely SelectKBest, which selects k features with the highest score [11] and ranks the features from the dataset based on the priority associated with the target variable. The framework of the diamond preprocessing dataset model is shown in Figure 1.

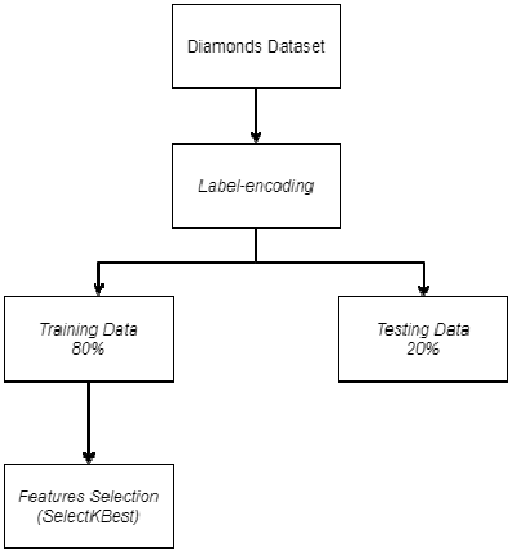

Fig. 1. Preprocessing of diamond dataset

D. Feature Correlation

Feature selection is important in many pattern recognition applications [12]. A good feature subset contains features that are correlated with predictions [13]. Irrelevant features can reduce the accuracy and quality of the model. From several features in the dataset, this research uses a library that has been provided by scikit-learn, namely SelectKBest. SelectKBest is the selection of features according to k with the highest score. This function can be called as select = SelectKBest() and performs fit_transform(X_train, y_train). In this study, the author uses the score_func parameter using f_regression, where the F value is between the labels/features for regression and k is 4, which will determine the ranking of the best four features. Table 1 and figure 2 show that carat, x, y, and z have an even distribution and the strongest correlation coefficient, price as a target for diamond prediction. So features other than in Table 1 are less relevant to price. The ranking results in coding are shown in Figure 2 and in Table 3.

SelectKBest for feature selection on all dataset

```
skb = SelectKBest(f_regression, k=4).fit(features, data['price'])
scores = skb.scores_
all_features = features.columns.values
sort_index = np.argsort(scores)[::-1]
rank = 1
ranked_features =[]
print("Ranking of features is ")
for x in sort_index:
    print(rank, ". Score  ", all_features[x], " is ", scores[x])
    ranked_features.append(all_features[x])
    rank += 1
print(all_features)
```

```
Ranking of features is
1 . Score   carat  is  304050.9059404779
2 . Score   x  is  193740.2790688721
3 . Score   y  is  160914.48180725984
4 . Score   z  is  154922.121056945
```

Fig. 2. Ranking Results in Coding

So it can be seen that using SelectKBest carat, x, y, z is quite relevant. Where x = length (mm) , y = width (mm) , z = depth (mm).

TABLE I. RANK 4 RESULTS FROM SELECTKBEST

Ranking of Features	Result	
	Features	*Score*
1	carat	304050.90
2	x	193740.28
3	y	160914.48
4	z	154922.12

From the results of the rank four values in Table 1, then the standard distribution graph can be seen in Figure 3.

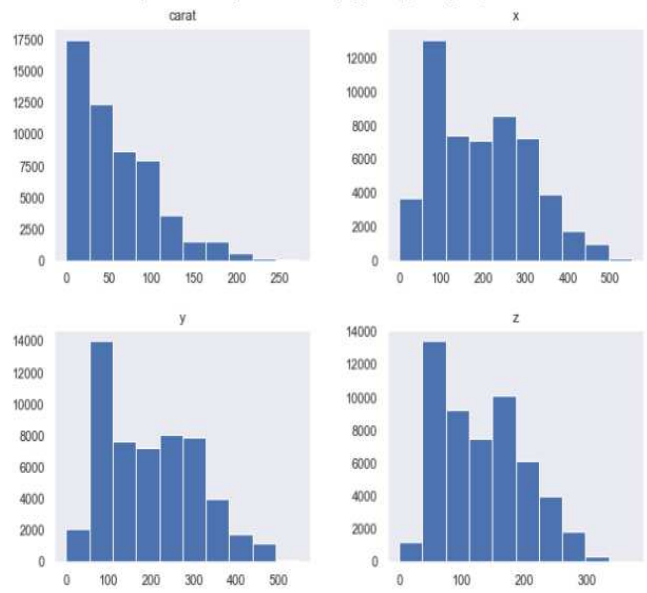

Fig. 3. Normal Distribution Modeling

E. *Modeling*

In this study, the authors chose machine learning regression because the value of diamond prices is continuous. Using the scikit-learn library, the two best regression models; k-Nearest Neighbors and LASSO, were selected for comparison.

1. *k-Nearest Neighbors* (k-NN)

k-Nearest Neighbors or k-NN is a method using a supervised algorithm. K-NN regression is an intuitive non-parametric method[14], with the feature of using the k approach to find a value that is close to the result by calculating the value of the proximity of the new case to the old case[15] where k is the number of nearest neighbors. The k-NN parameters of the scikit-learn modeling are set in Table 2.

TABLE II. PARAMETER K-NEIGHBORS REGRESSOR SCIKIT-LEARN

Parameter	Value	
	Value	Description
n_neighbors	Default = 5	Number of neighbors to use by default for neighbors queries.
weights	{'uniform', 'distance'} Default= 'uniform'	Weight used in the prediction.
metric	Default=	Distance metrics
	'minkowski'	like Euclidean, Minkowski.

2. *Least Absolute Shrinkage and Selection Operator* (LASSO)

The Least Absolute Shrinkage and Selection Operator (LASSO) is based on the concept of minimizing the standard mean square error by the sum of the absolute values of the regression coefficients[16]. Whereas LASSO will minimize the sum of the remaining squares on the sum of the absolute values of the coefficients smaller than the constant[17], the general form of this formula is in equation 1.

$$\hat{\alpha}, \hat{\beta} = \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2\right\} \quad (1)$$

Where:
(xi,yi) = data i = 1, 2, …, N
xi = predictor variable , (xi1,xi2,… xip)T
yi = response
The LASSO parameters of the scikit-learn modeling are set out in Table 3:

TABLE III. PARAMETER LASSO SCIKIT-LEARN

Parameter	Value	
	Value	Description
alpha	Default= 1.0	Constant that multiplies the L1 term. Defaults to 1.0. alpha = 0 is equivalent to an ordinary least square

III. RESULTS AND DISCUSSION

In this study, we use Root Mean Squared Error (RMSE) to examine the model's error rate and R-squared (R2) to measure and give the percentage of the total variation in the independent variable affecting the dependent variable. RMSE listed in [18], the best value is 0, and the worst value is + ∞

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (2)$$

Where :
i = variable i
m = number of data points that are not lost
Xi = actual observation time series
Yi = approximate time series

In addition, practitioners tend to look at R2, "Coefficient of multiples of determination," to assess fit according to[19]

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{RSS}{TSS} \quad (3)$$

Where :
y = prediction
y = average
yi = observed value of Y
RSS = regression sum of squares
TSS = total sum of squares

There are two metrics most commonly used to measure the accuracy of continuous variables, and Root Mean Squared Error (RMSE) and R-squared (R2). Table 3 shows the RMSE and R2 of the model that has been implemented. The k-NN and LASSO models were tested with different values of k and alpha. The difference in values is taken randomly, and tuning is done on the data to decide the best k and alpha values.

In k-NN, the best k value for the algorithm depends on the data that has been cleaned. In this study, it will be calculated with data that has been cleaned and is ready to be processed. To determine the k nearest neighbors, use neighbors and KneighborsRegressor provided by the scikit-learn library. In this test, to determine the k nearest neighbors using the neighbors and KNeighborsRegressor provided by the scikit-learn library, it is set in the range 1 to 30. Applying the Elbow method helps to select the optimal cluster from the results of this k-NN tuning. So it can be concluded from Table 4 and Figure 3 that k = 11 with Euclidean metric is the best result because, seen from the score results, take the value of k at the "elbow" that is the point after starting to decrease in value linearly. The results of the best k value tuning are in Table 4.

TABLE IV. BEST K VALUE TUNING RESULT

K	Score (RMSE)
1	1235.081
2	1089.058
3	1028.402
…	…
…	…
10	926.636
11	922.614
12	918.565
...	…
…	…
28	901.7793
29	901.5459
30	901.8806

The best k value can be described by a graph from the tuning results, as shown in Figure 4.

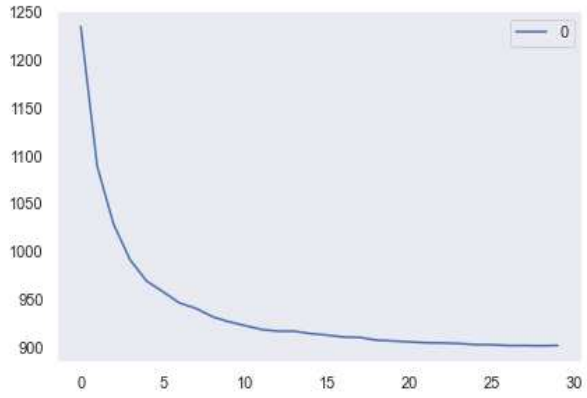

Fig. 4. Plotting of k . Tuning Results

The calculation of RMSE and R2_Score on k-NN is divided into 2, namely calculations with k values from 1 to 5, which can be seen in Table 5, and calculations with the best k values for tuning results, namely k=11, as shown in Table 5.

TABLE V. RMSE AND R2_SCORE WITH K VALUES 1-5

K	Result	
	RMSE	***R2_Score***
1	1235.08	0.834
2	1091.47	0.8703
3	1031.13	0.8843
4	995.26	0.8922
5	973.31	0.8969

TABLE VI. RMSE AND R2_SCORE WITH THE BEST K

K	Result	
	RMSE	***R2_Score***
11	926.06	0.9066

In LASSO, regression analysis with optimal alpha is required. The parameter used in this LASSO is alpha. In this study, GridSearchCV was used to find the optimal alpha by using the grid of alpha-value that the author had determined. The parameters used are as follows:
- estimator using lasso
- param_grid {'alpha':[5,10,20,30,35,40,45,50,55,100]}
cv (cross-validation) is 10
After using the parameters, proceed with the best_params_ attribute to issue the best alpha result from the specified param_grid. The results of this study, the optimal alpha result is 5, so in the experiment using alpha = 5. The RMSE and R2_Score values of param_grid are in Table 7.

TABLE VII. RMSE AND R2_SCORE OF PARAM_GRID

alpha	Result	
	RMSE	***R2_Score***
5	1049.59	0.8801
10	1050.14	0.88
20	1051.54	0.8796
30	1053.35	0.8792
40	1055.56	0.8787
45	1056.79	0.8784
50	1058.16	0.8781
55	1059.63	0.8778
100	1071.38	0.875

Comparing RMSE and R2 in both models, k-NN gives better results than LASSO. k-NN gives a minor RMSE result of 926.07 and the highest R2 of 0.9066 or 90.66%. The results of RMSE and R2 from the two models can be seen in Table 8.

TABLE VIII. RMSE AND R2 RESULTS FROM THE TWO MODELS

Model	The best result		
	Parameter	***RMSE***	R^2
k-NN	k = 11	926.06	0.9066
LASSO	alpha = 5	1049.59	0.8801

The form of visualization of the comparison is shown in Figure 5.

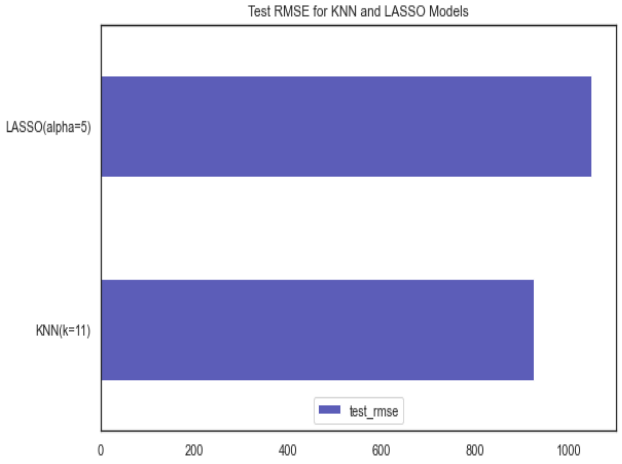

Fig. 5. RMSE comparison of KNN and LASSO models

ACKNOWLEDGMENT

Thanks to Faculty of Computer Science, Universitas AMIKOM Yogyakarta who helped in this study.

IV. CONCLUSION

The comparison results of RMSE and R2 on the k-NN and LASSO methods show that the k-NN method gives better results than the LASSO method. k-NN gives the minor RMSE result of 926.07 and the highest R2 of 0.9066 or 90.66%.

However, setting the parameter value affects the accuracy results. In addition, the selected features: carat, x, y,z, are not sufficient to predict the price of diamonds due to various factors, such as the quality of diamonds and reactions from social media that affect investment. Therefore, to get the best results from the model, the dataset should constantly be updated and added data.

REFERENCES

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

[2] T. Evans, "Diamonds," *Contemporary Physics*, vol. 17, no. 1, pp. 45–70, 1976, doi: 10.1080/00107517608210841.

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

[4] D. Bzdok, M. Krzywinski, and N. Altman, "Points of significance: Machine learning: Supervised methods," *Nature Methods*, vol. 15, no. 1, pp. 5–6, 2018, doi: 10.1038/nmeth.4551.

[5] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, "Data preprocessing in predictive data mining," *Knowledge Engineering Review*, vol. 34, 2019, doi: 10.1017/S026988891800036X.

[6] G. Y. Lee, L. Alzamil, B. Doskenov, and A. Termehchy, "A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.07127

[7] W. E. Winkler and / U S Bureau, "Data Cleaning Methods," 2003.

[8] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

[9] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *International Journal of ...*, vol. 1, no. 2, pp. 1–7, 2006, doi: 10.1080/02331931003692557.

[10] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, 2021, doi: 10.1080/00401706.2021.1921037.

[11] T. Desyani, A. Saifudin, and Y. Yulianti, "Feature Selection Based on Naive Bayes for Caesarean Section Prediction," in *IOP Conference Series: Materials Science and Engineering*, Aug. 2020, vol. 879, no. 1. doi: 10.1088/1757-899X/879/1/012091.

[12] P. and V. D. and K. C. Haindl Michal and Somol, "Feature Selection Based on Mutual Correlation," in *Progress in Pattern Recognition, Image Analysis and Applications*, 2006, pp. 569–577.

[13] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper," 1999. [Online]. Available: www.aaai.org

[14] Armando Teixeira-Pinto, "Machine Learning for Biostatistics," 2021.

[15] K. S. Y. Pande, D. G. H. Divayana, and G. Indrawan, "Comparative analysis of naïve bayes and knn on prediction of forex price movements for gbp/usd currency at time frame daily," in *Journal of Physics: Conference Series*, Mar. 2021, vol. 1810, no. 1. doi: 10.1088/1742-6596/1810/1/012012.

[16] J. Lv, M. Pawlak, and U. D. Annakkage, "Prediction of the transient stability boundary using the lasso," *IEEE Transactions on Power Systems*, vol. 28, no.1,pp. 281–288, 2013, doi: 10.1109/TPWRS.2012.2197763.

[17] R. Tibshiranit, "Regression Shrinkage and Selection via the Lasso," 1996.

[18] 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 Computer Science*, vol. 7, pp. 1–24, 2021, doi: 10.7717/PEERJ-CS.623.

[19] M. S. Lewis-Beck and A. Skalaban, "The /?-Squared: Some Straight Talk Downloaded from," 2015. [Online]. Available: http://pan.oxfordjournals.org/