Indonesian Community on Optimization and Computer Application
Vol. 1, No. 1, 2023, pp. xx-xx
ISSN: XXXX-XXXX, DOI:

Prediction Of Used Car Prices Using K-Nearest Neighbour, Random Forest And Adaptive Boosting Algorithm

Tiara Lailatul Nikmah[1*], Risma Maulidya Syafei[2], Rini Muzayanah[3], Asharinnisa Salsabila[4]
[1,2,3] Universitas Negeri Semarang, Semarang, Indonesia
* Corresponding Author

ARTICLE INFO

Article history:

Received July 28, 2021
Revised August 28, 2021
Accepted September 28, 2021

Keywords:

Price prediction;
Used cars;
K-nearest neighbor;
Random forest;
AdaBoost.

ABSTRACT

In the midst of busy society and high lifestyle, there are now many car offerings with advanced features. The more sophisticated a car is, the price increases. This makes people prefer to buy a used car with specifications that are still suitable for use. Therefore , used car entrepreneurs try to provide prices that are in accordance with the quality of the car. In order for the price of the used car offered to be in accordance with the market and not make used car entrepreneurs suffer losses, it is necessary to predict the right and accurate price. This study aims to help used car business owners to determine the appropriate car price using 3 algorithms, namely K-nearest neighbor, Random Forest and AdaBoost. The novelty of this study is the improvement in the accuracy of the prediction model of a single model. The results of this study are that the algorithm that has the best performance is Random Forest. This is shown by the smallest MSE and RMSE values among others. The MSE value is 117.142273 and the RMSE value is below 1 which is 0.342261.

Cite Article:

Tiara Lailatul Nikmah[1*], Risma Maulidya Syafei[2], Rini Muzayanah[3], Asharinnisa Salsabila[4], "Prediction Of Used Car Prices Using K-Nearest Neighbour, Random Forest And Adaptive Boosting Algorithm" *Indonesian Community on Optimization and Computer Application*, vol. 1, no. 1, pp. xx-xx, 2022, doi: .

1. INTRODUCTION

The increasing busyness of the community makes the development of technology more rapid because it has to adapt to the needs of the community. One of the technologies that are a necessity for people in the midst of busy schedules is cars. Now many cars with advanced features are introduced, thus making the price of cars increase [1]. Therefore, many people choose to buy used cars that are still suitable for use [2]. A large number of used car showrooms shows that people's interest in used cars is very high, and this certainly makes this business increase [3]. Used car transactions further pushed the used car market to its peak, and a number of problems slowly began to arise, such as the absence of a unified standard for assessing used car assets [4]. On the other hand, the problem that is often faced by used car business people is the right car pricing. Used car transactions are much more complex than other commodity transactions, since the selling price is influenced not only by the basic features of the car itself, such as brand, power, and structure, but also by the condition of the car [5]. Used car business owners need to make price estimates to minimize the risk of cars that are more expensive than needed [6], [7]. With the proliferation of the number of private cars and the advancement of the used car market, used cars should be the top priority of buyers. The price of a used car is an important aspect of a successful transaction for both buyers and sellers. Predicting the price of a used car is an interesting and much-needed issue to deal with [8]. Because the quality of the car and the price of the car affect the decision to buy a used car [9]. If the price and quality do not match there is a high probability

that the buyer will not so buy. Then it is necessary to analyze the right predictions regarding the price of used cars based on their specifications.Given the variety of factors affecting price variations, price prediction is always a difficult task [10]. The type of fuel used in cars as well as fuel consumption per mile greatly affects the price of the car because there are frequent changes in the price of fuel [11]. The product quality of a used car is a benchmark for consumers in assessing the feasibility of a car to buy. One strategy for predicting the price of an item is to use Machine Learning [12]. The machine learning model used to predict car prices as accurately as possible based on features uses 3 algorithms, namely K-nearest neighbour, Random Forest and Boosting algorithm. Then the research results will be compared with a single model of each algorithm used. The novelty of this study is the improvement in the accuracy of the prediction model of a single model.

2. METHODS

The method carried out in this study went through several stages, namely Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation. The stages of this method are shown in Figure 1. The data collection procedure is carried out at this stage. The data used is from Kaggle and is a secondary data type. After data collection, a preprocessing stage is carried out which consists of the stages of processing, cleaning, and data analysis [13]. Data preprocessing includes various stages, including encoding and standardization. Categorical data will be converted into numerical data by the process of enoding [14]. Encoding is the process of transforming information, such as text and images, from its original representation into an output format according to established rule [15]. The purpose of encoding is to try and use low-dimensional irregular features for representation by compressing high-dimensional features [16]. It is very important to filter out high-dimensional data before entering the learning process. The get_dummies function of the Pandas package is used during the encoding process of the category feature.

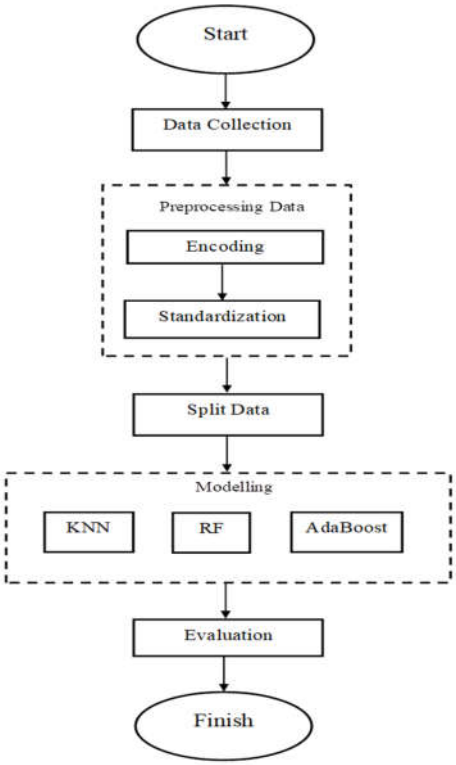

Fig. 1. Research method

Standardization is used to equalize the scale of data [17]. The ability to integrate different systems and processes requires standardization [18]. The importance of standardization because features that are relatively the same size and close to normal distribution can make machine learning algorithms more effective [19]. The StandardScaler function of the Python library is used during the data standardization process [20]. By eliminating scaling unit averages and fluctuations, the standard scaler is a preprocessing method that will normalize features [21]. After the preprocessing and data cleansing phases are completed,

the data will be divided 90:10 between the training data and the test data. Using the train_test_split function. This data sharing tries to prevent overfitting on the grounds that more training data will train the model more, make the model more robust and improve accuracy [22]. Create a used car price prediction model using the K-Nearest Neighbors (KNN), Random Forest, and Adaptive Boosting (AdaBoost) methods. KNN is an algorithm that estimates the value of each new piece of data using "feature similarity" [23]. By selecting a number of k from the nearest neighbor, KNN compared the distance between the two samples [24]. KNN has the advantage of being easy to understand and implement, strong against noisy training sample data, effective with large training sample data, and strong consistency. Random Forest is a classifier that improves the predictive accuracy of a data set with the average yield of many decision trees applied to different subsets of the data set [25]. Random Forest is one of the algorithms that uses bagging techniques. The bagging technique is trained by sampling with the replacement method. This method consists of several decision tree algorithms, whose features and data sharing are randomly selected [26]. This technique can handle huge volumes of data, data noise and missing value

AdaBoost is the third algorithm used. AdaBoost is one such algorithm that utilizes boosting techniques [27]. This technique works by creating a model from the training data. Then he developed a second model to correct the error in the first model. Until the training data is accurately predicted or the maximum number of models can still be added. This approach, which integrates several models of weak learners to form a strong ensemble learner [28]. The model will determine whether the observations made are accurate at each stage. The weights for each instance in the training data are the same. The wrong model is then given a higher weight to advance to the next round. The model is repeated until it reaches the required accuracy.The model will go through an evaluation step to determine how each model is performing. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used in the evaluation step. This evaluation aims to find the method with the minimum prediction error value and the best prediction result [29]. The model's ability to estimate prices is measured using metrics. The difference between the average of the actual values and the projected values is calculated by the MSE. The MSE value can be obtained from Equation (1).

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{Y'_i})^{\mathbf{2}} \tag{1}$$

RMSE (Root Mean Square Error) is the square root of the MSE. The RMSE value is used to describe the error rate of the model data used The square root of MSE is known as RMSE (Root Mean Square Error). The error rate of the model data used is described by the RMSE number [30]. The model accuracy value increases with decreasing RMSE. This metric's advantage is that it penalizes significant errors more severely, allowing for the possibility of more precision in some circumstances while avoiding the need of absolute value retrieval, which is undesirable in many mathematical operations [31]. The RMSE value is calculated using Equation (2).

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{Y_i} - \boldsymbol{Y'_i})^{\mathbf{2}}} \tag{2}$$

3. RESULTS AND DISCUSSION

This study conducted a used car price prediction using a model of 3 machine learning algorithms, namely K-nearest neighbour, Random Forest, and AdaBoost Algorithm. From these three algorithms, the model will be made as accurate as possible, that is, the model with the smallest possible error value. It will then evaluate the performance of each algorithm and determine which algorithm gives the best prediction results and which has the smallest prediction error value.

Table 1. Dataset attributes

Attribute Name	Type Data	Description
Unnamed: 0	Numerical	line number
Car Brands	Categorical	Car name
Pattern	categorical	Car models
Price	Numerical	used car selling price
Model Year	Numerical	year of production of the car
Place	Categorical	Locations where cars are sold or available for purchase
Fuel	Categorical	car fuel
Driven (Kms)	Numerical	number of Kilometers that the car has travelled
Tooth	Categorical	car gear transmission (Automatic/Manual)
Possession	Numerical	Whether the ownership is First Hand, Second Hand or other
EMI (monthly)	Numerical	monthly instalments are given to the buyer if buying the car

The data used comes from Kaggle from the Used Car dataset which amounts to 5918 data with 11 attributes. Of the 11 attributes, there are numeric and categorical attributes. An explanation of the data type and the definition of each attribute can be seen in Table 1. The data preprocessing stage is carried out, namely data cleaning from missing values and outliers. Outlier lifting is carried out by the IQR method by creating a lower limit and an upper limit. IQR calculations are found in Equations (3) and (4).

$$\text{Lower limit} = Q1 - 1.5 * IQR \quad (3)$$

$$\text{Upper limit} = Q3 + 1.5 * IQR \quad (4)$$

Perform the process of encoding category features with one-hot-encoding techniques. This encoding process is carried out with the get_dummies feature. Its function is to substitute categorical data values into numerical data. The category variables in this dataset are 'Car_Brand', 'Model', 'Location', 'Fuel' and 'Gear'. The reason for using this technique is that machine learning models cannot process categorical data, so it is necessary to convert categorical data into numerical data. The results of the encoding process are shown in Table 2.

Table 2. Categorical feature encoding results

Price	Model_Year	Driven	EMI	Car_Brand_Chevrolet	Car_Brand_Datsun	...	Fuel_Petrol + CNG	Fuel_Petrol + LPG	Gear_Manual	
350199	2011	20979	7790	0	0	…	1	0	0	1
306399	2011	19662	6816	0	0	…	1	0	0	1
208699	2015	11256	4642	0	0	…	1	0	0	0
249699	2012	28434	5554	0	0	…	1	0	0	1
240599	2011	31119	5352	0	0	…	1	0	0	1
...	...	...	...	...	...	…	...	...	...	...
523799	2014	46849	11652	0	0	…	1	0	0	1
295499	2013	68484	6573	0	0	…	1	0	0	1
405399	2015	59222	9018	0	0	…	1	0	0	1
348399	2019	30782	7750	0	0	…	1	0	0	1
551699	2018	67132	12272	0	0	…	1	0	0	1

Then the data standardization stage is carried out. StandardScaler performs the feature standardization process by subtracting the mean (the average value) and then dividing it by the standard deviation to shift the distribution. The StandardScaler returns a distribution with a standard deviation equal to 1 and a mean equal to 0. The results of the standardization are shown in Table 3.

Table 3. Results of standardization of numerical features

Model_Year	Driven	EMI
-1.359.962	0.096482	-1.013.887
-1.008.029	0.799182	-1.002.568
-0.304161	1.048.087	-0.320936
0.399707	0.781558	0.510561
-1.359.962	-0.123051	-1.463.480

After applying the 3 algorithms above, the prediction results were obtained that the AdaBoost algorithm had the closest results to the correct values. This is shown in Table 4.

Table 4. Model prediction evaluation table

prediksi_KNN	prediksi_RF	prediksi_AdaBoost
295723.5	304049.0	313039.8

Measurements of how well the model performs are also calculated from MSE and RMSE values. The results of the evaluation with MSE metrics are shown in Table 5 and Figure 2.

Table 5. Evaluate RMSE metrics

algorithm	train	Test
KNN	1163897.489566	1743989.054443
RF	79.249852	117.142273
AdaBoost	803786.49694	772898.050706

Table 6. Evaluate

algorithm	train	test
KNN	34.1159	41.7611
RF	0.281514	0.342261

AdaBoost	28.3511	27.801

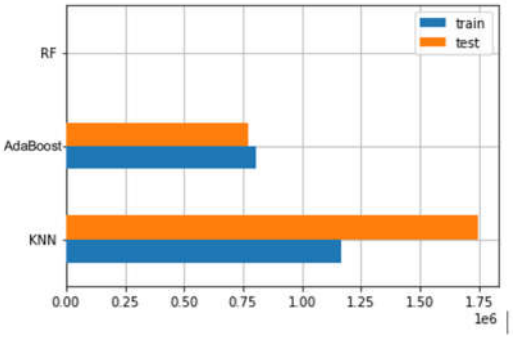

Fig. 2. MSE metric plot

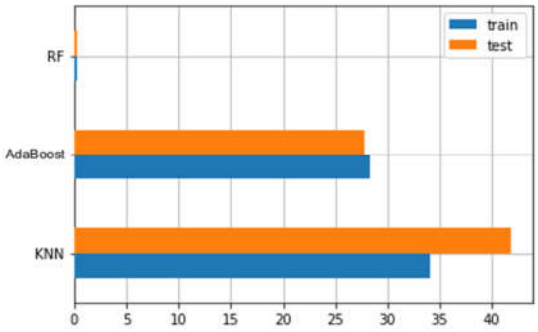

Fig. 3. MSE metric plot

Then also carried out evaluation calculations with RMSE metrics. The evaluation results of the RSME metrics are shown in Table 6 and Figure 3.From the calculation results of the MSE and RMSE metrics, the smallest error values were obtained in the Random Forest algorithm which had an MSE value of 117.142273 and RMSE 0.342261.

4. CONCLUSION

This research conducts used car price predictions to help buyers and sellers of used cars in determining the appropriate price. Many factors affect the price of a used car, one of which is the age of the car, the length of use, the number of times the user changes and others that must be calculated correctly. The Exploratory Data Analysis stage is also carried out to understand the relationship between categorical features. After the modelling and evaluation stage 3 machine learning models were used, namely KNN, Random Forest and AdaBoost. The algorithm that performs best is Random Forest. This is shown by the smallest MSE and RMSE values among others. The MSE value is 117.142273 and the RMSE value is below 1 which is 0.342261.

DECLARATION

Author Contribution

The method carried out in this study went through several stages, namely Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation.

Funding

Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER, grant number XXX” and “The APC was funded by XXX”.

Acknowledgments

In this section, you can acknowledge any support given which is not covered by the author con-tribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflict of Interest

Declare conflicts of interest or state “The authors declare no conflict of interest.”

REFERENCES

[1] A. Amalia, M. Radhi, S. H. Sinurat, D. R. H. Sitompul, and E. Indra, “Prediksi harga mobil menggunakan algoritma regressi dengan hyper-parameter tuning,” J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA), vol. 4, no. 2, pp. 28–32, 2022.
[2] C. Febrianita, M. Longgom, and N. Amalita, “Faktor–faktor yang mempengaruhi perilaku konsumen memilih mobil bekas merk toyota menggunakan analisis faktor,” J. Mat. UNP, vol. 2, no. 2, pp. 5–9, 2019.
[3] K. Bambang, Kurinawati, and H. F. Pardede, “Prediksi harga mobil bekas dengan machine learning,” Syntax Lit. J. Ilm. Indones., vol. 6, no. 5, p. 6, 2021.

[4] B. Cui, Z. Ye, H. Zhao, Z. Renqing, L. Meng, and Y. Yang, "Used car price prediction based on the iterative framework of XGBoost+LightGBM," Electron., vol. 11, no. 18, 2022.
[5] E. Liu, J. Li, A. Zheng, H. Liu, and T. Jiang, "Research on the prediction model of the used car price in view of the PSO-GRA-BP neural network," Sustain., vol. 14, no. 15, 2022.
[6] P. Susanti and K. Sussolaikah, "Penerapan metode regresi linear untuk memprediksi harga jual mobil bekas yaris dan jazz pada wilayah DKI Jakarta," vol. 7, no. 2, pp. 133–144, 2022.
[7] D. R. Das Adhikary, R. Sahu, and S. Pragyna Panda, "Prediction of used car prices using machine learning," in Biologically Inspired Techniques in Many Criteria Decision Making, 2022, pp. 131–140.
[8] A. Yadav, E. Kumar, and P. K. Yadav, "Object detection and used car price predicting analysis system (UCPAS) using machine learning technique," Linguist. Cult. Rev., vol. 5, no. S2, pp. 1131–1147, 2021.
[9] S. Wandi and H. Abaharis, "Pengaruh kualitas produk, promosi dan harga terhadap keputusan pembelian mobil bekas merek avanza di kota Padang," OSF Prepr., pp. 1–12, 2020.
[10] M. Ahtesham and J. Zulfiqar, "Used car price prediction with pyspark," in Digital Technologies and Applications, 2022, pp. 169–179.
[11] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car price prediction using machine learning techniques," TEM J., vol. 8, no. 1, pp. 113–118, 2019.
[12] S. K. Satapathy, R. Vala, and S. Virpariya, "An automated car price prediction system using effective machine learning techniques," in 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), May 2022, pp. 402–408.
[13] T. L. Nikmah, M. Z. Ammar, Y. R. Allatif, R. M. P. Husna, P. A. Kurniasari, and A. S. Bahri, "Comparison of LSTM , SVM , and naive bayes for classifying sexual harassment tweets," J. Soft Comput. Explor., vol. 3, no. 2, pp. 131–137, 2022.
[14] M. Al-Omari, M. Rawashdeh, F. Qutaishat, M. Alshira'H, and N. Ababneh, "An intelligent tree-based intrusion detection model for cyber security," J. Netw. Syst. Manag., vol. 29, no. 2, pp. 1–18, 2021.
[15] J. Lu et al., "Distributed information encoding and decoding using self-organized spatial patterns," Patterns, vol. 3, no. 10, p. 100590, Oct. 2022.
[16] Z. Yang, S. Lai, X. Hong, Y. Shi, Y. Cheng, and C. Qing, "DFAEN: Double-order knowledge fusion and attentional encoding network for texture recognition," Expert Syst. Appl., vol. 209, no. January, p. 118223, 2022.
[17] M. Roswell, J. Dushoff, and R. Winfree, "A conceptual guide to measuring species diversity," Oikos, vol. 130, no. 3, pp. 321–338, 2021.
[18] Y. Lu, X. Xu, and L. Wang, "Smart manufacturing process and system automation – A critical review of the standards and envisioned scenarios," J. Manuf. Syst., vol. 56, pp. 312–325, Jul. 2020.
[19] D. T. Wiranata et al., "Kecepatan tinggi turbin angin menggunakan machine learning dengan pendekatan support vector regression (SVR)," J. Tek. MESIN, vol. 9, no. 2, pp. 181–190, 2021.
[20] A. Ramdan, N. Widyasono, and H. Mubarok, "Prediksi jaringan TOR dan VPN menggunakan algoritma k-nearest neighbour pada trafik darknet," J. Sist. Cerdas, vol. 05, no. 01, pp. 21–35, 2022.
[21] V. Riandaru Prasetyo, M. Mercifia, A. Averina, L. Suntoyo, and Budiarjo, "Prediksi rating film pada website IMBD menggunakan metode neural network," J. Ilm. NERO, vol. 7, no. 1, pp. 1–8, 2022.
[22] B. Tenggehi, I. Palupi, "Prediksi perubahan kondisi uptrend dan downtrend pada pasar saham dengan menggunakan model artificial neural network ann," eProceedings, vol. 9, no. 3, pp. 2084–2093, 2022.
[23] P. Theerthagiri, I. J. Jacob, A. U. Ruby, and Y. Vamsidhar, "Prediction of COVID-19 possibilities using KNN classification algorithm," Int. J. Curr. Res. Rev., vol. 13, no. 6, pp. 156–164, 2021.
[24] S. Du and J. Li, "Parallel processing of improved KNN text classification algorithm based on hadoop," 2019 7th Int. Conf. Information, Commun. Networks, pp. 167–170, 2019.
[25] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House price prediction using random forest machine learning technique," Procedia Comput. Sci., vol. 199, pp. 806–813, 2022.
[26] X. Huang, H. Wang, W. Xue, A. Ullah, and S. Xiang, "A combined machine learning model for the prediction of time- temperature-transformation diagrams of high-alloy steels," J. Alloys Compd., vol. 823, p. 153694, 2020.
[27] A. Shahraki, M. Abbasi, and Ø. Haugen, "Engineering applications of artificial intelligence boosting algorithms for network intrusion detection : A comparative evaluation of Real AdaBoost , Gentle AdaBoost and Modest AdaBoost," Eng. Appl. Artif. Intell., vol. 94, no. July, p. 103770, 2020.
[28] M. Y. Chia, Y. F. Huang, C. H. Koo, and K. F. Fung, "Recent advances in evapotranspiration estimation using artificial intelligence approaches with a focus on hybridization techniques a review," 2020.
[29] M. G. Uddin, S. Nash, M. T. M. Diganta, A. Rahman, and A. I. Olbert, "Robust machine learning algorithms for predicting coastal water quality index," J. Environ. Manage., vol. 321, p. 115923, 2022.
[30] S. Ameer et al., "Comparative analysis of machine learning techniques for predicting air quality in smart cities," IEEE Access, vol. 7, pp. 128325–128338, 2019.
[31] G. K. Kamalam and S. Anitha, "Cloud-IoT secured prediction system for processing and analysis of healthcare data using machine learning techniques," Adv. Healthc. Syst. Empower. Physicians with IoT-Enabled Technol., pp. 137–172, 2022.