Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA), 3 (2) (2023) 81-89

Published by: LLDIKTI WILAYAH 1

Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA)

Journal homepage: www.lldikti1.ristekdikti.go.id/jurnal/index.php/jumpa

Random forest and decision tree algorithms for car price prediction

Purwa Hasan Putra[1], Azanuddin[2], Bister Purba[3], Yulia Agustina Dalimunthe[4]

[1,2,3,4]Jurusan Teknik Komputer dan Informatika, Politeknik Negeri Medan, Medan, Indonesia

Article Info

Article history:

Received: Mar 13, 2023
Revised: Apr 03, 2023
Accepted: Apr 21, 2023

Keywords:

Car price prediction;
Classification;
Data mining;
Decision tree;
Machine learning;
Random forest.

ABSTRACT

At this time in the era of cars that use renewable energy fuels such as electric cars which are highly supported by the government so that it has an impact on used cars based on these problems an analysis is needed. Determining whether or not the price of buying or selling a used car is appropriate is one of the obstacles faced by the community in making decisions when buying or selling a car or vehicle. Therefore, most people choose an alternative by buying a used car that is still good and usable. One way to make price predictions is to use the Machine Learning method. In this study the authors used random forest and decision tree methods to predict car prices. The results of the research on car price prediction analysis using the random forest and decision tree methods have different percentage results. Where using the random forest method there is an accuracy: 72.13% whereas with the analysis of the decision tree method accuracy: 67.21%. So it can be concluded that the Random Forest method has better analytical accuracy than the Decision Tree method.

Corresponding Author:

Purwa Hasan Putra,
Jurusan Teknik Komputer dan Informatika,
Politeknik Negeri Medan,
Jl. Almamater No.1, Padang Bulan, Kec. Medan Baru, Kota Medan, Sumatera Utara 20155, Indonesia.
Email: pputra@polmed.ac.id

INTRODUCTION

At this time in the era of cars that use renewable energy fuels such as electric cars which are highly supported by the government so that it has an impact on used cars based on these problems an analysis is needed. Determining whether or not the price to buy or sell a used car is one of the obstacles faced by the community in making decisions when buying or selling a car or Vehicle (Hasibuan et al., 2022). The price of a used car is influenced by several factors related to the car itself, such as the type of car, model, edition, year of production, transmission, fuel, engine capacity and mileage. Prices also fluctuate and competition is high among used car sellers, a tool is needed to predict used car prices accurately and quickly (Kriswantara & Sadikin, 2022).

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as OLX, Carsome and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market ("Used Cars Price Prediction Using Supervised Learning Techniques," 2019). Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features (Chandak et al., 2019).

Different websites have different algorithms to generate th retail price of the used cars, and hence there isn't a unified algorithm for determining the price. By training statistical models for predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website. The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy (Gajera et al., n.d.).

Along with the growing level of activity and business, cars have now become one of the needs of society. On the other hand, many advanced features are introduced to new cars, so the price of new cars increases significantly (East-West University et al., n.d.). Therefore, most people choose an alternative by buying a used car that is still good and usable. One way to make price predictions is to use the Machine Learning method (Amalia et al., 2022).

There are several algorithms for classification in data mining, including K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Support Vector Machines (SVM), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) (Sotarjua & Santoso, 2022). In this study the authors used the Random Forest and Decision Tree methods. Random Forest and Decision Tree are data mining algorithm methods that are used to classify a dataset (Pamuji & Ramadhan, 2021).

Random Forest can improve accuracy because there is a random selection in generating child nodes for each node (the node above it) and the classification results of each tree are accumulated, then the classification results that appear the most often are selected. (Saadah & Salsabila, 2021).

The Decision Tree is one of the most popular supervised learning-based methods for classification. It is an iterative top-down based approach (Zhang, 2021). A decision tree comprises a root node, decision nodes, and leaf nodes. The root node represents the most important attribute of the dataset used for obtaining the best prediction (Dutta et al., 2020).

This study aims to analyze the selling price of used cars by applying machine learning. Where currently encouraging the government to use cars that use environmentally friendly fuel or electricity by giving discounts on every purchase, so that it has an impact on car sales, especially used cars. The author applies machine learning using random forest and decision tree methods for analysis in order to get the best accuracy results.

In this study the authors used the https://www.kaggle.com/ dataset about predicting car prices. In the data set, it has been determined which variables are significant in predicting car prices. How well that variable describes the price of a car.

Based on previous research and problems that exist in the community, the authors conducted this study aimed at analyzing the Random Forest and Decision Tree methods to see the best accuracy results in predicting car prices. The benefits are expected to help the community in choosing a vehicle based on the variables from the data used.

RESEARCH METHOD

In this research two main research approaches were carried out, namely the qualitative approach and the quantitative approach. A qualitative approach is used to analyze the literature review regarding the variables that have an influence on used car prices. While the quantitative approach is a research method used to examine a particular population or sample of the dataset used to predict car prices.

The research methodology carried out consisted of 4 stages data understanding, data procesing, analysis random forest and decision tree and model evaluation. The following are theories related to the following research.

Machine learning.

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that can make decisions based on data. With a number of training data, the ML model is optimized on this data to produce a good predictive model, and can be used in future data. (Kriswantara & Sadikin, 2022).

Machine learning allows in data classification, this application recognizes patterns in data either with training or without training (Hasan Putra et al., 2022). Machine learning allows humans to program computers so that machines can recognize patterns or learn from what is put into them (Putra et al., 2022).

Random forest.

Random Forest (RF) is an algorithm that uses a recursive binary split method to reach the final node in a tree structure based on classification and regression trees (Prasad et al., 2006) (Sabri et al., n.d.) (Smarra et al., 2020). The Random Forest algorithm shows several advantages including being able to produce relatively low errors, good performance in classification, being able to handle large amounts of training data efficiently, and an effective method for estimating missing data (Wu et al., 2008). The Random Forest generates many independent trees with subsets selected randomly via bootstrap from the training sample and from the input variables at each node (Pamuji & Ramadhan, 2021).

The Random Forest (RF) method is a method that can improve accuracy results, because in generating child nodes for each node it is done randomly (Yang et al., 2008). This method is used to build a decision tree consisting of root nodes, internal nodes, and leaf nodes by taking attributes and data randomly according to the applicable provisions (Sarker et al., 2020). The root node is the final node that is located at the very top, or commonly referred to as the root of the decision tree (Song & Ying, 2015). Internal nodes are branching ends, where these nodes have at least two outputs and only one input. While the leaf node or terminal node is the last node that has only one input and has no output. The decision tree begins by calculating the entropy value as a determinant of the level of impurity of the attribute and obtaining value information. To calculate the entropy value, use the formula as in equation 1, while the information gain value uses equation 2 (Dutta et al., 2020).

$$Entropy\ (Y) = -\sum_i p\ (c|Y) log_2\ p(c|Y) \quad (1)$$

Where Y is the set of cases and p(c|Y) is the proportion of Y values to class c:

$$= Entropy\ (Y) = -\sum_{vt\ Values\ (\alpha)} \frac{Y_v}{Y_\alpha} Entopy\ (Y_v)\ ) \quad (2)$$

Where Values(a) is all possible values in case set a. Yv is a subclass of Y with class v which is related to class a. Yes are all values that correspond to a

Decision tree.

Decision Tree is an algorithm commonly used for decision makingv (Jijo & Abdulazeez, 2021) (Priyam et al., 2013). The Decision Tree will look for solutions to problems by making criteria as nodes that are interconnected to form a tree-like structure (Mohandoss et al., 2021). Decision tree is a predictive model for a decision using a hierarchical or tree structure (Papageorgiou & Stylios, 2006). Each tree has branches, branches represent an attribute that must be fulfilled to go to the next branch until it ends in a leaf (Guo et al., 2019). The concept of data in the Decision Tree is data expressed in the form of a table consisting of attributes and records (Pamuji & Ramadhan, 2021).

RESULTS AND ANALYSIS

In this research two main research approaches were carried out, namely the qualitative approach and the quantitative approach. A qualitative approach is used to analyze the literature review regarding the variables that have an influence on used car prices. While the quantitative approach is a research method used to examine a particular population or sample of the dataset used to predict car prices.

The research methodology carried out consisted of 4 stages, which can be seen as follows:

Figure 1. Research Methodology

Data understanding.

This stage is the data collection stage, analyzing the data to understand the data to be used, identifying problems by understanding the substance in the data and looking for interesting things in the data. The data in this study were obtained from the Kaggle dataset.

The data collection method to obtain the data source used is the secondary data collection method obtained from the kaggle.com source with a total of 205 rows and 9 columns of data. Table 1 is a display of some samples from the dataset:

Row No.	carbody	prediction(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	car_ID	symboling
1	hatchback	hatchback	0.051	0.440	0.418	0.048	0.044	10	0
2	sedan	hatchback	0.026	0.483	0.417	0.025	0.050	11	2
3	sedan	sedan	0.002	0.045	0.800	0.120	0.033	16	0
4	sedan	sedan	0.032	0.340	0.354	0.049	0.226	17	0
5	sedan	sedan	0.001	0.180	0.743	0.074	0.002	21	0
6	hatchback	hatchback	0.006	0.796	0.177	0.010	0.010	22	1
7	hatchback	hatchback	0.007	0.601	0.359	0.018	0.016	24	1
8	hatchback	sedan	0.001	0.208	0.723	0.066	0.002	25	1
9	wagon	wagon	0.000	0.034	0.175	0.790	0.001	29	-1
10	sedan	sedan	0.000	0.293	0.466	0.240	0.001	36	0
11	hatchback	hatchback	0.021	0.491	0.391	0.032	0.065	38	0
12	hatchback	hatchback	0.011	0.501	0.383	0.032	0.073	39	0
13	sedan	hatchback	0.021	0.592	0.299	0.018	0.070	43	1
14	sedan	sedan	0.000	0.045	0.783	0.123	0.048	49	0
15	sedan	hardtop	0.187	0.249	0.213	0.012	0.339	50	0
16	hatchback	hatchback	0.006	0.810	0.161	0.009	0.014	52	1
17	sedan	sedan	0.001	0.148	0.773	0.074	0.003	54	1
18	hatchback	hatchback	0.013	0.948	0.030	0.002	0.008	56	3
19	wagon	wagon	0.000	0.008	0.353	0.568	0.071	69	-1
20	sedan	sedan	0.001	0.018	0.695	0.147	0.139	71	-1
21	convertible	hatchback	0.051	0.631	0.111	0.004	0.203	73	3
22	hatchback	hatchback	0.007	0.707	0.223	0.013	0.051	79	2

Figure 2. Sample Dataset Used

In figure 2, the author uses the https://www.kaggle.com/ dataset about predicting car prices. In the data set, it has been determined which variables are significant in predicting car prices. How well that variable describes the price of a car. The car price dataset consists of 9 attributes with a total of 205 rows of data.

As for the process of predicting car prices, the author uses the random forest and decision tree methods to analyze car price predictions.

Data preprocesing.

The next process before the algorithm model is made is data preprocessing. In this research, preprocessing techniques were used, namely: cleansing, data aggregation, checking for missing values.

Random forest.

In this test using a random forest, for testing the author uses a rapid miner by displaying random forest data and methods with the following results.

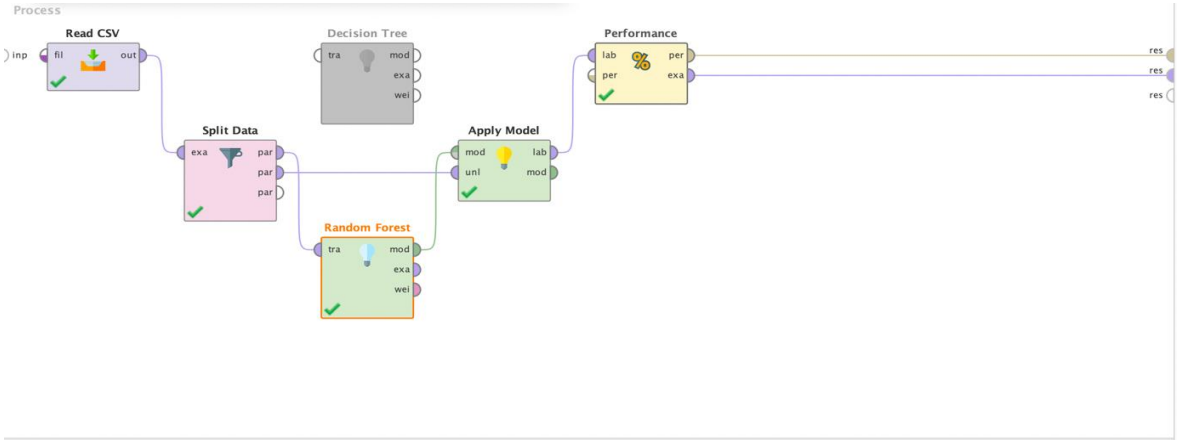

Figure 3. Random Forest Implementation Process

In Figure 3 is the process of the random forest method using rapid miner tools, where the author imports data with the read csv function, and splits data based on variables, selects the random forest method and then selects apply model, and selects performance to test the accuracy of applying the method so that the percentage end will be generated.

From the results of applying the random forest method, the authors get a pattern analysis of several car brands with several criteria from car predictions so that it can be seen that the results are shown as follows.

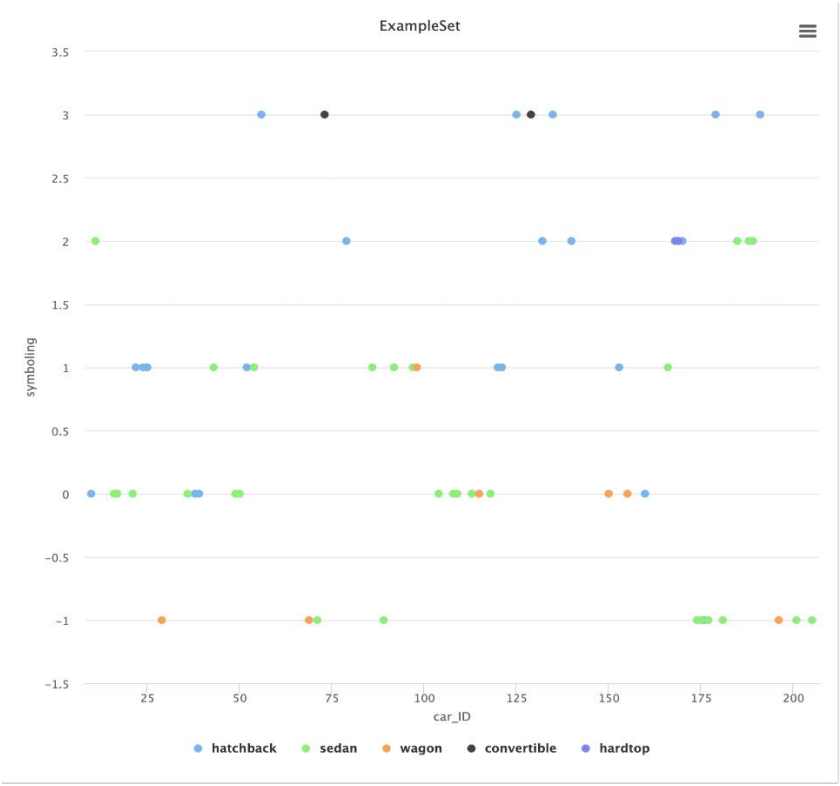

Figure 4. Graph of Random Forest Prediction Results

In Figure 4 the graphical results of the car price prediction process using the random forest method. The results of the predictions obtained several types of cars including hatchbacks, sedans, wagons, convertibles and hardtrops.

accuracy: 72.13%

	true convertible	true hatchback	true sedan	true wagon	true hardtop	class precision
pred. convertible	0	1	0	0	0	0.00%
pred. hatchback	1	15	4	0	2	68.18%
pred. sedan	0	5	24	2	0	77.42%
pred. wagon	0	0	0	5	0	100.00%
pred. hardtop	1	0	1	0	0	0.00%
class recall	0.00%	71.43%	82.76%	71.43%	0.00%	

Figure 5. Persentage Accuracy Random Forest

In Figure 5 is the percentage of accuracy of the calculation process which has 72.13%, from the results of this accuracy there are 5 types of cars that have a class percentage, such as predictions of hatchbacks with 68.18%, sedans 77.42%, wagons 100%, convertibles 0.00 % and hardtrops 0.00%

Performance vector.

Table1 below explains prediction results using the Random Forest method, it has an accuracy of 72.13% in predicting car prices. From the results of the accuracy, there is a matrix value for each type of car where the matrix value for the type of hatchback car has the highest value of 15.24 for the type of hatchback and sedan.

Table 1. Result Accuracy Method Random Forest

accuracy: 72.13% Confusion Matrix:					
True:	convertible		hatchback	sedan	wagon	hardtop
convertible:		0	0	0	0	0
hatchback:		1	15	4	0	2
sedan:	0	5	24	2	0	
wagon:	0	0	0	5	0	
hardtop:	1	0	1	0	0	

Decision tree.

In figure 6 the author applies the decision tree method for testing using rapid miner tools, where the author uses read csv to import the car price dataset, then splits the data based on the existing variables, then connects the filtered data with the decision tree, selects the apply model and connecting performance to produce accurate results from each dataset used.

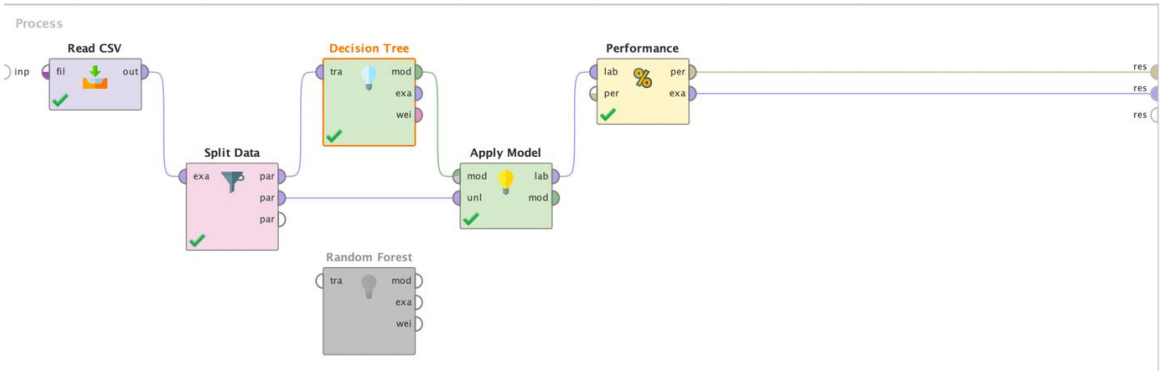

Figure 6. The Process of Implementing the Decision Tree

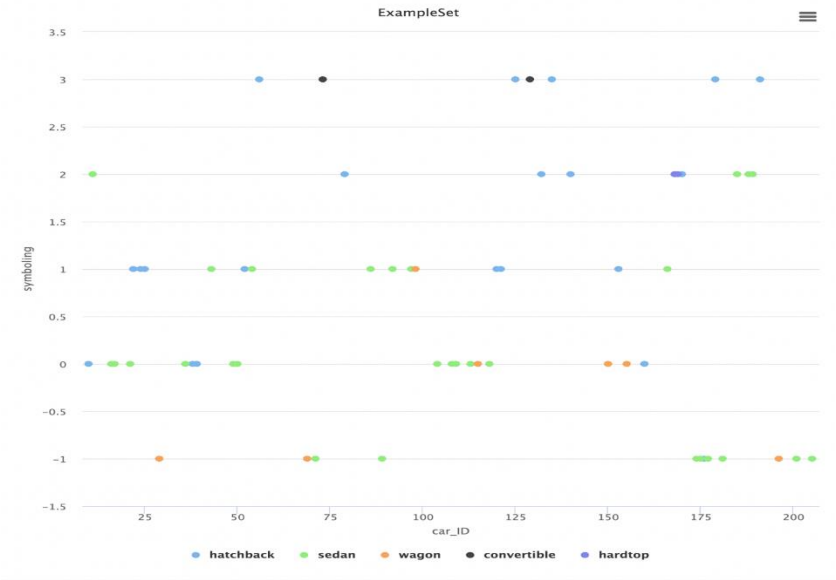

Figure 7. Graph Results of the Decision Tree Method

In figure 7 the results of the prediction graph using the decision tree, there are 5 types of hatchback, sedan, wagon, convertible, hardtop cars.

accuracy: 67.21%

	true convertible	true hatchback	true sedan	true wagon	true hardtop	class precision
pred. convertible	0	0	0	0	0	0.00%
pred. hatchback	1	17	7	0	2	62.96%
pred. sedan	0	4	19	2	0	76.00%
pred. wagon	0	0	3	5	0	62.50%
pred. hardtop	1	0	0	0	0	0.00%
class recall	0.00%	80.95%	65.52%	71.43%	0.00%	

Figure 8. Persentage Accuracy Decision Tree

In Figure 8 is the percentage of accuracy of the calculation process which has 67.12%, from the results of this accuracy there are 5 types of cars that have a class percentage, such as predictions of hatchbacks with 6796.18%, sedans 76.00%, wagons 62.50%, convertibles 0.00 % and hardtrops 0.00%.

Performance vector.

In table 2 the prediction results using the decision tree method have an accuracy of 67.21% in predicting car prices.From the results of the analysis of car price predictions using the random forest and decision tree methods, the percentage results are different. Where using the random forest method there is an accuracy of: 72.13% while the decision tree analysis method has an accuracy of: 67.21%. So it can be concluded that the Random Forest method has better accuracy analysis than the Decision Tree method.

Table 2. Result Accuracy Method Decision Tree

accuracy: 67.21%						
ConfusionMatrix:						
True:	convertible		hatchback	sedan	wagon	hardtop
convertible:		0	0	0	0	0
hatchback:		1	17	7	0	2
sedan:	0	4	19	2	0	

accuracy: 67.21%					
ConfusionMatrix:					
wagon:	0	0	3	5	0
hardtop:	1	0	0	0	0

CONCLUSION

At this time in the era of cars that use renewable energy fuels such as electric cars which are highly supported by the government so that it has an impact on used cars based on these problems an analysis is needed. Determining whether or not the price of buying or selling a used car is appropriate is one of the obstacles faced by the community in making decisions when buying or selling a car or vehicle. Therefore, most people choose an alternative by buying a used car that is still good and usable. One way to make price predictions is to use the Machine Learning method. In this study the authors used random forest and decision tree methods to predict car prices. From the results of the analysis of car price predictions using the random forest and decision tree methods, the percentage results are different. Where using the random forest method there is an accuracy: 72.13% whereas with the analysis of the decision tree method accuracy: 67.21%. So it can be concluded that the Random Forest method has better analytical accuracy than the Decision Tree method. In the future, it will be interesting to investigate car price predictions based on external factors, for example, other issues that are not related to the specifications of the car itself. At the moment it seems that current issues, economic conditions and others can affect the increase in car prices. Interesting to be one of the future research developments, especially in variable costs and methods that are in accordance with car selling price predictions.

REFERENCES

Amalia, A., Radhi, M., Sinurat, S. H., Sitompul, D. R. H., & Indra, E. (2022). Prediksi Harga Mobil Menggunakan Algoritma Regressi Dengan Hyper-Parameter Tuning. *Jurnal Sistem Informasi Dan Ilmu Komputer Prima(JUSIKOM PRIMA)*, *4*(2), 28–32. https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2479

Chandak, A., Ganorkar, P., Sharma, S., Bagmar, A., & Tiwari, S. (2019). Car Price Prediction Using Machine Learning. *International Journal of Computer Sciences and Engineering*, *7*(5), 444–450. https://doi.org/10.26438/ijcse/v7i5.444450

Dutta, K. K., Sunny, S. A., Victor, A., Nathu, A. G., Ayman Habib, M., & Parashar, D. (2020). Kannada alphabets recognition using decision tree and random forest models. *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, 534–541. https://doi.org/10.1109/ICISS49785.2020.9315972

East-West University, Institute of Electrical and Electronics Engineers, Institute of Electrical and Electronics Engineers. Bangladesh Section, & IEEE Robotics and Automation Society. Bangladesh Chapter. (n.d.). *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT 2019) : May 3-5, 2019, Dhaka, Bangladesh.*

Gajera, P., Gondaliya, A., & Kavathiya, J. (n.d.). OLD CAR PRICE PREDICTION WITH MACHINE LEARNING. In *International Research Journal of Modernization in Engineering Technology and Science* *www.irjmets.com @International Research Journal of Modernization in Engineering*.

Guo, Y., Zhou, Y., Hu, X., & Cheng, W. (2019). Research on recommendation of insurance products based on random forest. *Proceedings - 2019 International Conference on Machine Learning, Big Data and Business Intelligence, MLBDBI 2019*, 308–311. https://doi.org/10.1109/MLBDBI48998.2019.00069

Hasan Putra, P., Syahputra Novelan, M., & Rizki, M. (2022). Analysis K-Nearest Neighbor Method in Classification of Vegetable Quality Based on Color. *Journal of Applied Engineering and Technological Science*, *3*(2), 126–132.

Hasibuan, E., Informasi, S., Ilmu, F., Informasi, T., Gunadarma, U., Margonda, J., No, R., Cina, P., & Jawa, D. (2022). Implementasi Machine Learning untuk Prediksi Harga Mobil Bekas dengan Algoritma Regresi Linear berbasis Web. *Jurnal Ilmiah Komputasi*, *21*(4), 595–602. https://doi.org/10.32409/jikstik.21.4.3327

Jijo, B. T., & Abdulazeez, A. M. (2021). Classification based on decision tree algorithm for machine learning. *Evaluation*, *6*(7).

Kriswantara, B., & Sadikin, R. (2022). Used Car Price Prediction with Random Forest Regressor Model. *Journal of*

Information Systems, Informatics and Computing Issue Period, *6*(1), 40–49. https://doi.org/10.52362/jisicom.v6i1.752

Mohandoss, D. P., Shi, Y., & Suo, K. (2021). Outlier Prediction Using Random Forest Classifier. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference, CCWC 2021*, 27–33. https://doi.org/10.1109/CCWC51732.2021.9376077

Pamuji, F. Y., & Ramadhan, V. P. (2021). Komparasi Algoritma Random Forest dan Decision Tree untuk Memprediksi Keberhasilan Immunotheraphy. *Jurnal Teknologi Dan Manajemen Informatika*, *7*(1), 46–50. https://doi.org/10.26905/jtmi.v7i1.5982

Papageorgiou, E., & Stylios, C. (2006). A Combined Fuzzy Cognitive Map and Decision Trees Model for Medical Decision Making. *Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA*, 6117–6120. https://doi.org/10.1109/IEMBS.2006.260354

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*, 181–199.

Priyam, A., Abhijeeta, G. R., Rathee, A., & Srivastava, S. (2013). Comparative analysis of decision tree classification algorithms. *International Journal of Current Engineering and Technology*, *3*(2), 334–337.

Putra, P. H., Hasibuan, A., & Marpaung, E. A. (2022). *Analisis Klasifikasi Metode X-Means Pada Minat dan Bakat Anak Dimasa Pandemi*. *19*(2), 424–429.

Saadah, S., & Salsabila, H. (2021). Prediksi Harga Ponsel Menggunakan Metode Random Forest. *Jurnal Komputer Terapan*, *7*(1), 24–32.

Sabri, M. A., Yahyaouy, A., Tairi, H., El Beqqali, O., Benali, H., Jāmiʿat Sīdī Muḥammad ibn ʿAbd Allāh. Faculty of Sciences Dhar El Mahraz, IEEE Computer Society, & Institute of Electrical and Electronics Engineers. (n.d.). *2018 International Conference on Intelligent Systems and Computer Vision (ISCV) : April 2-4, 2018, Faculty of Sciences Dhar El Mahraz (FSDM), Fez, Morocco*.

Sarker, I. H., Colman, A., Han, J., Khan, A. I., Abushark, Y. B., & Salah, K. (2020). Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. *Mobile Networks and Applications*, *25*, 1151–1161.

Smarra, F., Di Girolamo, G. D., De Iuliis, V., Jain, A., Mangharam, R., & D'Innocenzo, A. (2020). Data-driven switching modeling for mpc using regression trees and random forests. *Nonlinear Analysis: Hybrid Systems*, *36*, 100882.

Song, Y.-Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130.

Sotarjua, L. M., & Santoso, D. B. (2022). Perbandingan Algoritma Knn, Decision Tree,* Dan Random* Forest Pada Data Imbalanced Class Untuk Klasifikasi Promosi Karyawan. ... *Informatika Sains Dan ...*, *7*(2), 192–200.

Used Cars Price Prediction using Supervised Learning Techniques. (2019). *International Journal of Engineering and Advanced Technology*, *9*(1S3), 216–223. https://doi.org/10.35940/ijeat.a1042.1291s319

Wu, L. C., Horng, J. T., Huang, H. Da, & Chen, W. L. (2008). Identifying discriminative amino acids within the hemagglutinin of human influenza A H5N1 virus using a decision tree. *IEEE Transactions on Information Technology in Biomedicine*, *12*(6), 689–695. https://doi.org/10.1109/TITB.2008.896871

Yang, B.-S., Di, X., & Han, T. (2008). Random forests classifier for machine fault diagnosis. *Journal of Mechanical Science and Technology*, *22*, 1716–1725.

Zhang, B. (2021). Tactical Decision System of Table Tennis Match based on C4.5 Decision Tree. *Proceedings - 2021 13th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2021*, 632–635. https://doi.org/10.1109/ICMTMA52658.2021.00146