JURNAL ILMIAH TEKNIK INDUSTRI
ISSN: 1412-6869 (Print), ISSN: 2460-4038 (Online)
Journal homepage: http://journals.ums.ac.id/index.php/jiti/index
doi: 10.23917/jiti.v21i2.18975

Data Mining Analytics Application for Estimating Used Car Price During the Covid-19 Pandemic in Indonesia

Bramantiyo Eko Putro[1a◆], Dwi Indrawati[1b]

Abstract. *Covid-19 has resulted in an increase in the people's need for vehicle ownership in order to avoid public transportation. People's purchasing power, on the other hand, has also weakened. Therefore, they prefer to purchase affordable cars, such as used cars. Moreover, the Luxury Goods Sales Tax (PPnBM) discounts were officially applied to the purchase of the new cars in March 2021. This study aims at estimating the price of used cars using several data mining algorithms, such as Random Forest, K-Nearest Neighbour (KNN), and Naïve Bayes. By employing the RapidMiner tool, this study was able to evaluate the attributes affecting car prices. From the experimental results, random forest producers have the highest accuracy of 95.46%. Then, this study figured out that brand, engine capacity, kilometres, colours, years, number of passengers, and transmissions are the most influential attributes to determine the estimation of the used car prices.*

Keywords: *data mining; estimation; used cars; random forest; k-nearest neighbour; naïve bayes.*

I. INTRODUCTION

Sales of four-wheeled vehicles before the pandemic had shown a positive trend during 2018. Car sales were able to break through a new psychological figure of 1.15 million units three years earlier, only in the range of 1 million units amid global economic stagnation. In addition, a fierce competition is again occurring in the family car segment with the presence of newcomers (Tamara, 2020). This increase was also seen in used car sales at the beginning of 2018, the number of used car sales increased dramatically, which was up to 90% from the previous year. The significance of this increase in demand is also evidenced by an increase in the price of used cars by around 10 percent and the proliferation of marketplaces to find used cars (Olga, 2019).

The Covid-19 outbreak has altered the mobility of developing countries such as Indonesia. As public transport is often crowded, particularly in developing countries, and it is difficult to maintain social distance under such conditions, public transport vehicles become potential hotspots for virus transmission (Abdullah et al., 2021). Ministry of Transportation Republic of Indonesia reported public transportation passengers to decline around 40 to 70 percent (Desk, 2020). The reduction is caused by people shifting to private vehicles such as cars (Abdullah et al., 2020), bicycles, or even walking (Habib et al., 2021). The index of the people's need for vehicle ownership has increased. People's purchasing power, on the other hand, has also weakened. Therefore, they prefer to purchase affordable cars, such as used cars. 54% of buyers prefer used cars to new cars during the Covid-19 outbreak according to BeliMobilGue.co.id and OLX Indonesia research report (Widodo et al., 2021).

As a result of the increase, the used car dealer business has become more competitive. Consequently, marketing strategy through marketing mix becomes vital for a business to survive (Syapsan, 2019). Price is an important element in the marketing mix to be able to win the competition. Price has become a strong factor influencing car purchasing decisions (Liao et al., 2016). Prices for cars, particularly the used ones, do not come straight from the manufacturer. Therefore, predicting car prices is a critical, important, and interesting issue (Pudaruth, 2014). To set the price, there is a lot of information that must be considered, from the brand, model,

[1] Department of Industrial Engineering, Faculty of Engineering, Universitas Suryakancana, Jl. Pasir Gede Raya, Cianjur, Jawa Barat, 43216.

[a] email: bramantiyo@unsur.ac.id
[b] email: dwindra2999@gmail.com
◆ corresponding author

Submited: 04-07-2022 Revised: 20-11-2022
Accepted: 01-12-2022

variant, year of purchase, year of the sale, engine capacity, transmission, number of passengers, kilometers, dimensions, and weight, even to the customers' color preferences (Pal et al., 2017; Pudaruth, 2014; Samruddhi & Ashok Kumar, 2020). In addition, research to estimate car prices is essential at this time. This is since the competition comes not only from fellow used car sellers, but also from new car sellers, once the Luxury Goods Sales Tax Incentive (PPnBM) of up to zero percent for the purchase of motorized vehicles was officially implemented in early March 2021 (Sera, 2021). The used car dealers begin to acknowledge the the impact of the PPnBM discount on new cars. Customers acknowledge the price difference between cars that are only 1-3 years of use as not being significantly different from new cars receiving PPnBM discounts. Consequently, sales of used cars that are only 1-3 years of use are declining.

Several analytical approaches can be used to estimate prices, one of which is data mining. Data mining uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases (Mirza, 2018). One of the six tasks in data mining is estimation (Larose & Larose, 2014). Estimation in a broad sense deals with continuously valued outcomes in which some input data will become a value for some unknown continuous attributes (B & G, 2013).

A previous study has predicted the used car prices. The results showed that the second method was easily influenced by the sample size. It is shown in Model 1 that the Regression method is better than the Random Forest method, proven by more stable results. However, the Regression model is not recommended for predicting the price of used cars due to the high sample requirements. Model 2 shows that both methods have a better predictive effect and are proportional to the increase in the sample size. Model 3 shows that the Random Forest model outperforms the regression model by five times with a prediction ability of 95.06%. This proves that, when compared to a simple model with few attributes, the Random Forest method is the most effective way to handle 11 complex models with a large number of attributes and sample (Chen et al., 2017). Another research compared the results of four different techniques in predicting used car prices consisting of linear regression, k-NN, and Naïve Bayes and Decision Tree. All techniques produced satisfactory performance, but the scoring accuracy was not as excellent as it may be due to several limitations such as a low record (Pudaruth, 2014).

Nabarun Pal et al. has predicted used car prices using machine learning Random Forest and Linear Regression (Pal et al., 2017). The author used big data containing the prices and attributes of over 370,000 used cars sold on the website across 40 brands with 20 attributes. The dataset was retrieved from the Kaggle dataset which is scraped from eBay-Kleinanzeigen. It was found that random forest regression performed better with 95,82% accuracy. The prediction model consists of price, kilometer, brand, and vehicle Type as the most relevant features. Another study conducted by Samruddhi and Kumar tried to predict the price of used cars using the KNN algorithm technique (Samruddhi & Ashok Kumar, 2020). The author used the KNN algorithm because it is suitable for small data set. The dataset used in this study was also taken from Kaggle. The data set contains 14 variables which include an unnamed serial number, Name, location, mileage, Fuel_Type, Engine transmission, Kilometers_Driven, Power, New_Price, Year, Seats, Owner_Type, Price. The study also resulted in better prediction than linear regression with 85% accuracy.

A recent study performed by Amik et al. tried to predict the price of used cars in Bangladesh. Used cars are popular in Bangladesh but there are not many online services accommodating the transaction between owners and buyers. This case has a similar condition to Indonesia. This study uses several machine learning techniques such as linear regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression, decision tree, random forest, and extreme gradient boosting. The dataset was scraped from https://bikroy.com/ contains 1209 instances and 10 features. The study results show that XGBoost has the best accuracy with 91,32% followed by Random Forest 90,14%.

Having informed by the previous works, this study, henceforth, compared several methods for estimating car prices based on the database on the Mobil123 website. The selection was based on the cases to be studied. The Random Forest method, one of the classifications and regression-based methods where there is a decision tree aggregation process (Primajaya & Sari, 2018), was the first method used in this study. The second method was K-Nearest Neighbors (KNN), a non-parametric method that can manage both classification and regression problems as one of the simplest ones of all machine learning algorithms. A sample is classified by estimating the majority vote of its neighbors, with the new object assigned to the class that is most common among its nearest neighbors (k being a positive integer, and typically small) (Triguero et al., 2019). Lastly, Naïve Bayes method was employed in this study. This method is a simple learning algorithm utilizing Bayes' rule, together with a strong assumption that the attributes are conditionally independent given to the class (Webb, 2010). In certain cases, utilizing probability is more effective than applying strict rules for classification. However, this is not the case with tree algorithm, which do not employ any rules (Putro & Saepurohman, 2020). Therefore, the purpose of this study is to determine which estimation model can provide the maximum level of accuracy when calculating the price of a used car. Furthermore, this study also seeks the most influential attributes in estimating used car prices. The results of this study were intended to be beneficial to car dealers who are affected by 0% PPnBM.

II. RESEARCH METHOD

Data Collection

The study employed the secondary data. Secondary data refer to information gathered by others for purposes other than the purpose of the current study (Sekaran & Bougie, 2016). This study obtained used cars data from the Mobil123 website (www.mobil123.com) which is listed until March 2021. The data population is 56,495 data, while the data used as samples in this study was 10,000 data. Data retrieval is done by the data scraping method or web scraping. This uses a software that simulates human browsing on the web to collect detailed information from different websites (Diouf et al., 2019). Web scraping's main objective is to extract information from one or more websites and convert it into simple formats like spreadsheets, databases, or CSV files (Diouf et al., 2019). This study utilized Parsehub as a web scraper tool. Sampling was carried out because the Parsehub free subscription can only extract about 200 pages of data in one run. Moreover, there were only 10,000 data records that could be processed with the tool during the data processing stage. The data collected include price, brand, model, variant, year of purchase, engine capacity, transmission, seat amount, kilometers or travel distance, and the color of the car.

Pre-processing

Many missing values, value distortion, misrecording, incorrect spelling, and insufficient sampling can all be found in raw data. Therefore, it is necessary to improve the quality by conducting data preparation (pre-processing). The pre-processing steps carried out in this study are:

1. Overcoming missing values

   Missing values are a persistent issue that wreaks havoc on data analysis methods (Larose & Larose, 2014). Several methods can be used to deal with missing values. This study eliminated records or fields from the data by purposely left them blank.
2. Outlier detection

   Outliers are extreme values that are opposite to the trend of the data (Masrofah & Putro, 2020). Outliers can lead information and data conclusions to deviate. It is because there are differences in mean and variance. Therefore, it is necessary to detect outliers.
3. Transformation

   The transformation stage is carried out so that the subsequent mining process will be more efficient, and the patterns discovered will be easier to understand (Han et al., 2012). Particularly in the KNN algorithm, the transformation is necessary to calculate the distance (closeness of location) of the training data set. Euclidean distance is the distance parameter used in the KNN algorithm that prefers numeric form (integer or real). Henceforth, attributes with nominal data such as brand, model, variant, transmission, and car color will be transformed.

4. Normalization

   Normalization is the process of scaling the attribute values of the data so that they can fall within a certain range (Han et al., 2012). Min-Max methods were used in this study by performing a linear transformation of raw data shown in equation (1) (Larose & Larose, 2014). This normalization still maintains the relationship between the actual data values.

$$\boldsymbol{X^{*}_{mm}} = \frac{\boldsymbol{X-\min(X)}}{\boldsymbol{range\ (X)}} = \frac{\boldsymbol{X-\min(X)}}{\boldsymbol{\max(X)-\min(X)}} \qquad \textbf{... (1)}$$

Data Processing

Data processing in this study are as follows:

1. Random Forest

   The algorithm stages to be followed when constructing a tree using a random forest are (Breiman, 2001; Wang et al., 2018):

   a. Performing random sampling of size n with the recovery on the cluster of data. This stage is the bootstrap stage.

   b. Employing the bootstrap example, the tree is constructed until it reaches its maximum size (without pruning). At each node, the disaggregation is done by selecting m explanatory attributes randomly, where m << p. The best disaggregation was selected from the m explanatory attributes. This stage is often called random feature selection.

   c. Repeating Step 1 and 2 in k times, to form a forest consisting of k trees. The response of observation was predicted by aggregating the predicted results of k trees.

2. Naïve Bayes

   The stages of the Naïve Bayes algorithm are (Zhang & Li, 2007):

   a. Preparing training data.

   b. Presenting each data as an n-dimensional vector, namely X=(x1,x2,x3,…….xn).

   c. Determining n as a description of the size made in the test from n attributes, namely A1, A2, A3…………., An.

   d. Determining m as a collection of categories, namely C1, C2, C3, ……..Cm.

   e. Given X test data whose category is unknown, the classifier will predict that X belongs to the category with the highest posterior probability based on the condition X.

   f. Naïve Bayes classifier marks the unknown test X to categorize C1 if and only if P(Ci|X)>P(Cj|X) for 1 ≤ j ≤ m, j ≠ i

   g. Then, we need to maximize P(Ci|X) shown in equation (2).

$$\boldsymbol{P(Ci|X)} = \frac{\boldsymbol{P(X|Ci).P(Ci)}}{\boldsymbol{P(X)}} \qquad \textbf{... (2)}$$

   h. Where x is the attribute value in sample X, the probability P(x1|Ci), P(x2|Ci),…… P(xn|Ci) can be estimated from the training data.

3. K-Nearest Neighbor (KNN)

   The steps of KNN are as follows (Kuang & Zhao, 2009):

   a. First, the data pre-processing phase is to initialize the labeled d-dimensional train data set as well as the test data set to be classified.

   b. Second, select one test point in the test data set and calculate the distances between it and each point in the train data set.

   c. The next phase is to sort the results of distances computation, and find out K smallest results according to the parameter K.

   d. The fourth step is to determine the class label of the test point by the election result of K points.

   e. Finally, select another point in test data set and go to step two repeatedly until the test data set is empty.

f. In general, to define the distance between two objects, x, and y, the Euclidean distance formula is used, shown in equation (2) (Masrofah & Putro, 2020):

$$\boldsymbol{D(i,j)} = \sqrt{\left(x_{1i}+x_{1j}\right)^2+\left(x_{2i}+x_{2j}\right)^2+\cdots+\left(x_{ki}+x_{kj}\right)^2} \ldots(3)$$

Model Evaluation

The model that has been formed needs to be evaluated before being implemented in the real world. The evaluation is intended to measure the level of performance of the model that will be generated. The measurement towards the model performance can be undertaken through several types of model evaluation, including Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Confusion Matrix.

III. Result and Discussion

Data Collection

Cars specifications were sold at mobil123.com as shown in Figure 1. Data were collected in March 2021 using the web scrapping method and the ParseHub software tool with the algorithm as shown in Figure 2. Used car data collected is 10,000 with 10 attributes.

Pre-processing

Missing value. The variant attributes and a quiet view in the model section are where the missing value in the used car data was generally found. This missing value can arise for one of two reasons: the seller may have overlooked it or may have purposefully left it blank. To overcome this, it is done manually by finding out the specifications of used cars from various resources on the internet. As a result, some missing values can be filled in, while the rest cannot due to the fact that the used car only has variant information. Therefore, deleting the data was carried out to overcome the issue of the missing value that cannot be filled in. The deletion of missing value data from 10,000 used car data resulted only in 9,624 of all columns with complete attributes.

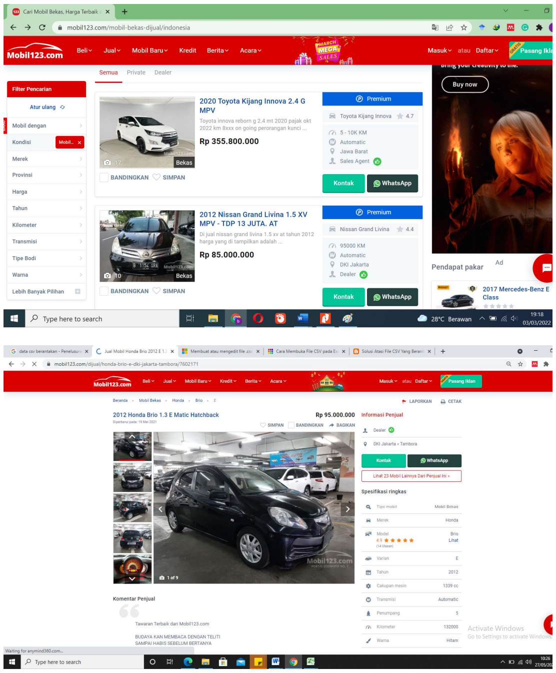
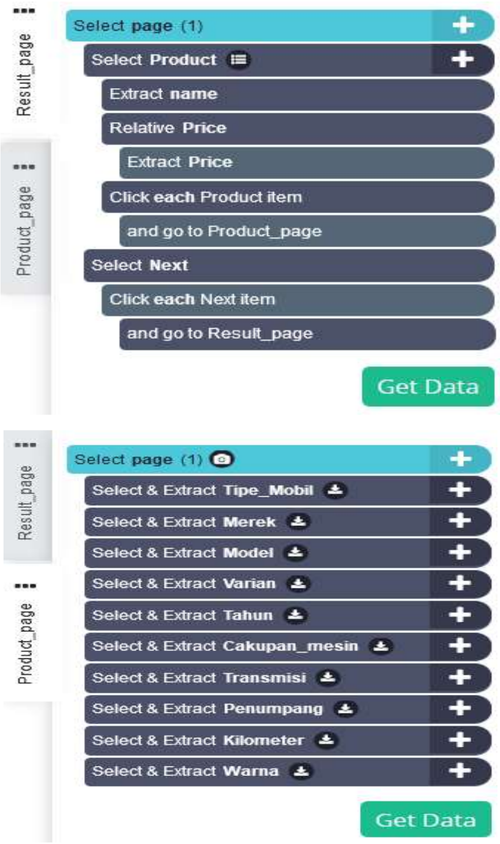

Mobil123.com
2020 Toyota Kijang Innova 2.4 G MPV
Rp 355.800.000
2012 Nissan Grand Livina 1.5 XV MPV - TDP 13 JUTA. AT
Rp 85.000.000
2012 Honda Brio 1.3 E Matic Hatchback
Rp 95.000.000
Kontak
WhatsApp

Figure 1. Webisite Mobil123.com


Result_page
Select page (1)
Select Product
Extract name
Relative Price
Extract Price
Click each Product item
and go to Product_page
Select Next
Click each Next item
and go to Result_page
Product_page
Get Data

Result_page
Select page (1)
Product_page
Select & Extract Tipe_Mobil
Select & Extract Merek
Select & Extract Model
Select & Extract Varian
Select & Extract Tahun
Select & Extract Cakupan_mesin
Select & Extract Transmisi
Select & Extract Penumpang
Select & Extract Kilometer
Select & Extract Warna
Get Data

Figure 2. ParseHub Algorithm

Transformation. Data transformation is carried out to change the original data (after eliminating the missing value) into data that is ready to be processed. The transformed data are data whose form is still in the nominal (letters) form. The process is carried out because data processing using KNN cannot use nominal data but the value must be an integer or real to calculate the Euclidean distances. Furthermore, Naïve Bayes data processing cannot use numeric labels; instead, they must be in nominal or categorical form. Data transformed in the form of number initialization are brand, model, variant, transmission, and color data. The price, on the other hand, is formed in the letter initialization.

Outlier detection. From the stage of searching and deleting outlier data, there are a total of 8401 data that can be used in the next stage.

Normalization. There are several sorts of attributes in the used car dataset to be processed. As a result, there is a wide range of values between these attributes. One of the reasons is that they use different units, such as prices and transmissions, which range from tens to hundreds of millions. When the attributes used have values with different ranges or scales, it can trigger poor model performance when performing data mining processes. The normalization process is needed to measure or scale the data of the attributes. Therefore, their values are in a smaller range of values. Normalization is done with the help of Rapid Miner with the normalized operator.

Data Processing

Random Forest. In the Random Forest algorithm, the setting for n-estimators (the number of trees) and the maximum depth in each forest is 10, which means that the study will produce 10 decision tree models with a maximum of 10 nodes. Since in this study an estimate was made, the parameter criterion chosen was least-square. This minimizes the squared distance between the average value in the node and the actual value, so that the less error is made possible. Performance measurement uses Root Mean Squared Error (RMSE) as a measure to evaluate the regression model by measuring the accuracy of the estimation results of a model. Based on 10 tree models made, a brand is the most important attribute in making price estimation decisions. This is because most brands appear as root nodes in 4 out of 10 tree models. The second important attribute is the range of machines that are root nodes in 3 out of 10 tree models. Furthermore, an important attribute in estimating the price of a used car is the kilometer which is the root node in 2 out of 10 tree models.

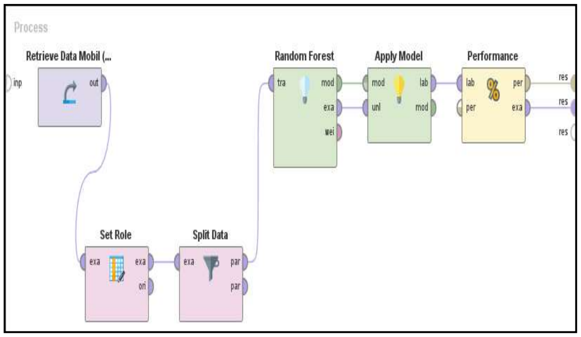


Figure 3. Random Forest Estimation

The fourth important attribute is the color that is the root node in 1 out of 10 tree models. After brand, engine capacity, kilometers, and color, year becomes another important attribute. The year often appears as an internal main node (parent node) in some trees and often becomes a child node again for other parent nodes. Furthermore, the transmission becomes a factor in the decision to estimate the price of a used car. In the study of used car price estimation using the random forest algorithm, the models and variants did not become nodes at all. It is widely known that Random Forest has a feature selection process that can select best features to increase model performance (Dewi & Chen, 2019). Because the models and variants in this study are diverse, the tree becomes too vast to be included in the tree model.

KNN. In the KNN algorithm, settings for k (nearest neighbor class) are made using the trial & error method to see which value of k produces the smallest error. The value of k is usually a small, positive, and odd integer (RapidMiner). Previous research predicted the price of used cars to determine the value of k = 1, 3, 5, and 10 and the best performance results were at k = 1 for cars with the Toyota brand, and k = 5 for cars with the Nissan brand (Pudaruth, 2014). Based on these references, in this study, the values of k that will be used are 1, 3, and 5. The "proximity" in this algorithm is defined in terms of distance so that the algorithm will calculate the value of the Euclidean distance. Similar to Random Forest the KNN algorithm on Rapid Miner will use the least square to quantify estimation performance, resulting in a Root Mean Squared Error (RMSE) value.

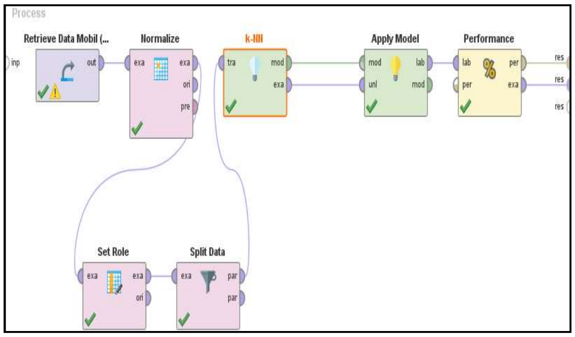


Figure 4. KNN Estimation

An important problem that arises in the use of the KNN method is the determination of the value of k. If the value of k given is correct, then the accuracy results can be more optimal and higher than the results of model accuracy using other algorithms. Whereas, if the value of k is less precise, it will be less accurate or inferior to those of other algorithms. Like the accuracy results in this study, the KNN model using k=1 and k=3 resulted in an accuracy of 91.13% and 63.58%, respectively, was greater than the accuracy using the Naive Bayes algorithm, which was 63.02%. However, the KNN model using k=5 produces an accuracy that is smaller than Naive Bayes, which is 53.55%. In this study, it can also be seen that the greater the value of k is, the greater the error occurs. Therefore, in this study, the shorter the neighboring class is, the better the accuracy is. Then, due to optimizing the data, it is necessary to carry out a normalization process in data processing, but the values removed from the model are also all in the interval -3 to 3 which makes it difficult to analyze the actual estimated price if it is not continued in the machine learning processing.

Naïve Bayes. In contrast to Random Forest and KNN, the performance setting to evaluate the estimation accuracy results in Naïve Bayes is to use a confusion matrix, because the label (price) is made in a categorical form (classification).

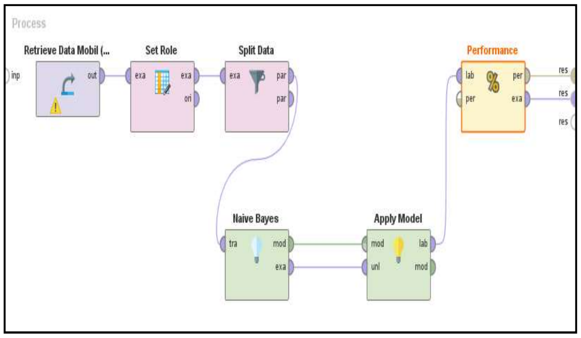


Figure 5. Naïve Bayes Estimation

Based on the results of data preparation and processing, the Naive Bayes algorithm cannot handle the numerical label data, such as prices. Therefore, price data must be initialized in nominal terms (alphabet A, B, C, ..., etc.). However, the variation in the price data in this study is very large. This is because the seller has complete control over the pricing of his car. Even though the make, model, variant, and the color of the car are the same, the prices they set can differ. For this reason, it will be difficult to initialize all data in the nominal form, considering that 10,000 data are processed. Therefore, the initialization of prices is done in the form of intervals. The results of data processing show that the model's estimated price is also in the form of initialization or category, which means it is also in the form of an interval. Therefore, it can be said that this method is not effective for estimating used car data because it does not produce a direct decision on the number use of a used car. Naïve Bayes also does not demonstrate a relationship between attributes. Because it assumes that each attribute is independent, whereas usually in a decision making, there is a correlation between attributes.

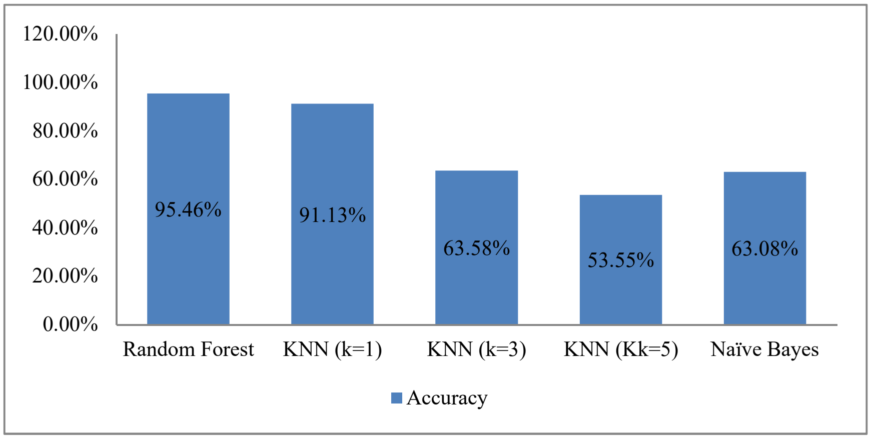


Figure 6. Comparison Graph of Accuracy of Random Forest, KNN, and Naïve Bayes Methods

After estimating the price of a used car based on data on the mobil123.com web using the random forest, KNN, and Naive Bayes methods, the next step is to compare the estimation results of the three methods used to find out which method is the best for estimating the price of a used car. Figure 6 shows a comparison graph of the accuracy of the random forest, KNN, and Naive Bayes methods.

Based on the accuracy results, the Random Forest algorithm generates a higher percentage of accuracy compared to the other algorithms. In addition, the use of this algorithm is more efficient in processing data in nominal and numeric forms. Therefore, transformation or normalization are not required. This algorithm also shows the decision-making steps for estimating prices as well as important attributes that influence price estimation decisions.

Discussion

Data processing to estimate the price of a used car using a random forest algorithm with n-tree 10 and a maximum depth tree of 10 produces 10 tree models with different root nodes and node sequences. To find out what the price of a used car is based on car attributes (brand, model, variant, year, engine coverage, transmission, color, kilometers, and passengers), an analysis process must be carried out on the 10 tree models. The value of the price that comes out more on the estimate in each tree is the answer. For example, if it is known that the car brand is Audi with model A4, variant 1.8 TFSI PI, purchased in 2013, is black, has a 1798 cc engine coverage with automatic transmission, can carry 5 passengers, and has been used for up to 45000 km, what is the estimated price of the car?

Based on observations on these 10 trees, 2 tree models estimate that a car with the Audi brand, model A4, variant 1.8 TFSI PI, purchased in 2013, is black, has a 1798 cc engine coverage with automatic transmission, can carry 5 passengers, and has been used up to 45000 km for Rp. 219,000,000, while 8 other tree models estimated the car at Rp. 275,000,000. Therefore, the decision on the price of the car is taken based on the mode, which is Rp. 275,000,000.

Based on 10 tree models made, a brand is the most important attribute in making price estimation decisions. This is because most brands appear as root nodes, namely in 4 out of 10 tree models. The reputation of European brands makes the selling value of their cars higher than cars from other countries. A previous study also found that people prefer brands from certain countries and the preference order differs between countries (Helveston et al., 2015), for example, the BMW, Mercedes-Benz, Audi, Land Rover Jaguar, Bentley, and Rolls Royce brands. A powerful brand inspires potential consumers to pay a premium price for a product. A premium price can be described as the amount or the level of utility that a customer is willing to pay for a brand, compared to other similar brands, and it can be either negative or positive (Ashraf et al., 2017). If people only buy, the company may be able to sell a large number of items yet not earn a profit. Market-based assets improve a firm's levels of inflowing cash in two ways: increasing volumes and setting higher prices (Bondesson, 2012). Moreover, based on the Ministry of Industry of the Republic of Indonesia (Kemenperin), 21 vehicles get a zero percent tax relaxation consisting of 6 brands. The six brands are Toyota, Daihatsu, Mitsubishi, Suzuki, Honda, SGMW (Wuling) (Keputusan Menteri Perindustrian Republik Indonesia Nomor 169 Tahun 2021, 2021). Therefore the used car price from these brands will be lower than the others.

The second important attribute in this study is the range of machines rooting nodes in 3 out of 10 tree models. Engine capacity, acceleration time, or maximum speed are related to the car's performance (Liao et al., 2016). It also affects the efficiency of fuel consumption used. The fact that cars with lower engine power use more fuel can be explained by the practice of curtailing engine power to profit from a lower motor vehicle tax (Meyer & Wessely, 2009). Fiscal policies have sought to counter these trends and internalize the negative externalities associated with increasing energy use and emissions from cars by incentivizing the purchase of lower-emitting cars (Rogan et al., 2011). Therefore this is reflected in the price of new and used cars. Another thing that makes this engine capacity very influential on car prices is the implementation of 0% PPnBM for cars with 1,500 cc engine capacity (Sera, 2021). Therefore, a used car seller with a 1500 cc engine capacity will indeed sell at a lower price than the standard selling price.

Furthermore, an important attribute in estimating the price of a used car is the kilometer rooting the node in 2 out of 10 tree models. Kilometers can be used as an indicator to find out whether the used car is still in good condition and worth purchasing. A large proportion of the used cars sold in 2020 belonged to the category of 50,000-80,000 kilometers (Mendiratta, 2021). The odometer showing 100,000 km indicates that the cars have degraded from its ideal performance. At that mileage, the engine has to be turned off, and some of the components have dirt on them and begin to wear out.

The distance-based approach is used as an assessment of emissions and emission comparisons between transport modes, it is measured either per kilometer or passenger kilometer (pkm) (Hagedorn & Sieg, 2019). The travel distance /kilometer has a significant relationship with the concentration of CO emissions, where the farther the distance is, the higher the concentration of CO is. CO emissions depend on the combustion of the fuel-air mixture and the presence of carbon content in the fuel (Ramalingam & Rajendran, 2019). Carbon deposits formed on the injector may cause operational problems, such as excessive smoke emissions, loss of power, poor fuel economy, degraded emissions, excessive engine noise, rough engine operation, and poor drivability (Suryantoro et al., 2016). These carbon deposits naturally cause self-ignition and accelerate engine damage, in addition to wasting fuel and reducing engine power. Therefore, it is obvious that kilometers will demonstrate how well a used car performs; the farther the distance/kilometer is, the lower the price is.

The fourth important attribute is the color, which is the root node in 1 out of 10 tree models. Color can affect resale prices both negatively and positively. Some colors are more desirable because they can disguise scratches, dings and dents, and dirt. Based on the research, the color of the car affects the incidence of accidents, although other aspects must still be considered (Newstead & D'Elia, 2007; Shin, 2013). Cars with popular colors such as white, black, and silver sell for higher prices and it is high rate protection on the used car market than similar cars in unique or less popular colors (Gong et al., 2018). According to a study by iSeeCars, yellow is the most valuable color in terms of highest resale value over three years. iSeeCars executive analyst Karl Brauer says that yellow may not be a widely desired car color, but there are quite few who want it, compared to the number of new yellow cars ordered. Yellow is one of the colors with the lowest vehicle share and it is most commonly associated with sports cars and other low-volume vehicles with good grades. The report examines 5.6 million new cars and 700,000 used cars purchased and sold between 2017 and 2020 to evaluate which colors to fix, damaged, or appear to be good in terms of car resale value. Krem, the second safest on the list, has depreciated only 22.8 percent in three years. Yellow, beige, orange, green, gray, red, blue, silver, white, black, purple, brown, and gold are the 13 colors listed in order. Yellow is best reported as the most popular color for SUVs, sedans, and pairs, while beige is the most popular one for pick-up trucks. When it comes to convertibles, red is preferable, whereas blue is great for minivans (KIRO, 2021).

After brand, engine capacity, kilometers, and color, year becomes the next important attribute. The year often appears as an internal main node (parent node) in some trees and often becomes a child node again for other parent nodes. The age of the car shows how long the car has been used by the previous owner. The older the car is, the lower the engine performance is. Cars that are less than 5 years of use commonly have a relatively safe resale price, considering the fact that they have not had any major repairs. While cars over 5 years of use will normally incur a fee for component replacement, lowering the overall cost of the car. Five to eight years old cars were observed to dominate the used car sales in Indonesia in 2020 (Mendiratta, 2021).

Furthermore, transmission becomes a factor in the decision to estimate the price of a used car. Manual transmission is more efficient in fuel consumption than automatic transmission, moreover, it is easy and cheaper to be manufactured (Zainuri et al., 2017). Therefore used cars price with manual transmission in the market has a cheaper price than automatic transmission. Finally, what influences the price estimation decision is the number of seats/passengers. In some cases, the number of seats is a consideration of customers to buy a car, especially for family cars. According to Ken Research report, multi-Purpose Vehicles or MPVs were observed to dominate the used car market based on sales volume as they are suitable for large-sized families (Mendiratta, 2021). However, it is not a major factor since sports cars that only carry two passengers can be rather costly.

Managerial Implications

Based on the results of the findings in this study, several policy implications can be recommended, including the following priorities that can be provided as input to the interested parties:

1. The results of this study can be used by mobile123.com to inform website development by including features to estimate prices for potential sellers based on variables that impact price estimation decisions. This development can help customers in estimating the resale price of their cars without trouble or confusion, and it will be undoubtedly bring value to mobil123.com by increasing the convenience of the website users.
2. Based on the results of the study, the brand is the first attribute that influences the decision to estimate the price of a used car. Therefore, the recommended implication is to pay attention to the brand when buying a car because some brands have the best resale price or after sale price. The car's brand has a significant impact on how quickly or slowly its value depreciates. Due to the well-known reputation and high quality, shrinkage tends to be slow. This will help reduce losses in reselling process.
3. The results of the study indicate that machine coverage is the second important attribute in price estimation decisions. If a person wish to sell a car with 1500 cc engine capacity purchased before the 0% PPnBM was applied, the seller must set a price even lower.
4. The seller must pay attention to the car to ensure that it has a good safety rating and history, as well as to the mileage to ensure that it is not excessive. This is since the mileage or kilometers have significant impact on the buyer's appraisal of the car condition. To maintain a reasonable selling price, attempts are taken to keep the kilometers at a value of 100,000 km.

5. The results of this study can also be used as guide for the banks when determining a car's depreciation based on its characteristics.

IV. CONCLUSION

Based on the research results of used car price estimation using 3 algorithms, the random forest has the highest accuracy value of 95.46%. KNN (k=5), on the other hand, was the algorithm with the lowest accuracy at 53.55%. Naive Bayes is not recommended for estimating prices because of its inability to handle output classes with numeric values. With the advantage of random forest that can build a price estimation decision model for each car, the choice of using a random forest to estimate the price of a used car becomes more effective, efficient, and convenient. Based on the results of a random forest analysis with 10 trees and a maximum depth of 10, the car brand is the most essential attribute that can affect the estimated price of a used car. While the models and variants in this study have no effect or are considered irrelevant, they can add to the complexity of the used car price estimation decision tree model and are automatically filtered out of the node by the random forest feature. Therefore, the most relevant attributes used for this estimation are brand, engine capacity, kilometers, color, year, number of passengers, and transmission.

The flaws and inadequacies cannot be overlooked when conducting this research. Therefore, researchers need to provide suggestions for further research in order to improve it. The following are some of the suggestions: The random forest algorithm can be used to reduce the amount of variance in the study into how the role of models and variants affect the pricing of used cars. RapidMiner's weight by important tree algorithm, which calculates the weight of each attribute's importance, cannot be used to estimate with a numeric label like price. Therefore, it is advised that the future research be conducted using software that can support search weights such as Python or WEKA. Mobil123.com, henceforth, can develop features on its website to estimate the price of a used car using the model developed in this study. Developing an automated interactive system with a used car repository until an estimated price estimate is available. This allows users to find out the price of the car using the recommendation feature.

REFERENCES

Abdullah, M., Ali, N., Javid, M. A., Dias, C., & Campisi, T. (2021). Public transport versus solo travel mode choices during the COVID-19 pandemic: Self-reported evidence from a developing country. *Transportation Engineering, 5*, 100078. https://doi.org/10.1016/J.TRENG.2021.100078

Abdullah, M., Dias, C., Muley, D., & Shahin, M. (2020). Exploring the impacts of COVID-19 on travel behavior and mode preferences. *Transportation Research Interdisciplinary Perspectives, 8*, 100255. https://doi.org/10.1016/j.trip.2020.100255

Ashraf, S. F., Li, C., & Mehmood, B. (2017). A Study of Premium Price Brands with Special Reference to Willingness of Customer to Pay. *International Journal of Academic Research in Business and Social Sciences, 7*(7). https://doi.org/10.6007/ijarbss/v7-i7/3126

B, R., & G, S. (2013). Application of Data Mining In Marketing. *International Journal of Computer Science and Network, 2*(5), 41–46.

Bondesson, N. (2012). Brand Image Antecedents of Loyalty and Price Premium in Business Markets. *Business and Management Research, 1*(1). https://doi.org/10.5430/bmr.v1n1p32

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, C., Hao, L., & Xu, C. (2017). *Comparative analysis of used car price evaluation models*. AIP Conference Proceedings, 1839(May). https://doi.org/10.1063/1.4982530

Desk, N. (2020). *Public transportation ridership drops by 70 percent*. City - The Jakarta Post. The Jakarta Post, 1. https://www.thejakartapost.com/news/2020/03/20/public-transportation-passenger-numbers-drop-up-to-70-percent.html

Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on features selection for regression analysis. *International Journal of Innovative Computing, Information and Control, 15*(6), 2027–2037. https://doi.org/10.24507/ijicic.15.06.2027

Diouf, R., Sarr, E. N., Sall, O., Birregah, B., Bousso, M., & Mbaye, S. N. (2019). *Web Scraping: State-of-the-Art and Areas of Application*. Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019, 6040–6042. https://doi.org/10.1109/BigData47090.2019.9005594

Gong, J., Peng, L., & Li, J. (2018). *A Study on the Factors Affecting the Value of Used Cars in Panzhihua Region*. Proceedings of the 2nd International Forum on Management, Education and Information Technology Application (IFMEITA 2017), 99–104. https://doi.org/10.2991/ifmeita-17.2018.17

Habib, M. A., Asif, M., & Anik, H. (2021). Impacts of COVID-19 on Transport Modes and Mobility Behavior: Analysis of Public Discourse in Twitter. *Transportation Research Record, 0*(0), 1–14. https://doi.org/10.1177/03611981211029926

Hagedorn, T., & Sieg, G. (2019). Emissions and external environmental costs from the perspective of differing travel purposes. *Sustainability, 11*(24). https://doi.org/10.3390/SU11247233

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. In Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5

Helveston, J. P., Liu, Y., Feit, E. M. D., Fuchs, E., Klampfl, E., & Michalek, J. J. (2015). Will subsidies drive electric vehicle adoption? Measuring consumer preferences in the U.S. and China. *Transportation Research Part A: Policy and Practice, 73*, 96–112. https://doi.org/10.1016/j.tra.2015.01.002

KIRO, 7 News Staff. (2021). *Study: Car color can affect resale value – KIRO 7 News Seattle*. KIRO 7, 1. https://www.kiro7.com/news/local/study-car-color-can-affect-resale-value/6QISJKYNPRERDF5CP7GZVGOGKM/

Kuang, Q., & Zhao, L. (2009). *A practical GPU based KNN algorithm*. International Symposium on Computer Science and Computational Technology (ISCSCT), 7(3), 151–155.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining.* Second Edition Wiley Series on Methods and Applications in Data Mining.

Liao, F., Molin, E., & Wee, B. van. (2016). Consumer preferences for electric vehicles: a literature review. *Transport Reviews, 37*(3), 252–275. https://doi.org/10.1080/01441647.2016.1230794

Masrofah, I., & Putro, B. E. (2020). *Clustering of the water characteristics of the Cirata reservoir using the k-means clustering method*. The 5Th International Conference on Industrial, Mechanical, Electrical, and Chemical Engineering 2019 (Icimece 2019), 2217, 030010. https://doi.org/10.1063/5.0000672

Mendiratta, A. (2021). *Indonesia Used Car Market Outlook to 2025 – By Market Structure (Organized & Unorganized), By Type of Car (MPVs, Hatchbacks, SUVs & Others), By Brand (Toyota, Honda, Daihatsu, Suzuki & Others), By Vehicle Age, By Mileage, By Customer Age and By Region* (DK. https://www.kenresearch.com/automotive-transportation-and-warehousing/automotive-and-automotive-components/indonesia-used-car-market-outlook-to-2025/412166-100.html#details

Meyer, I., & Wessely, S. (2009). Fuel efficiency of the Austrian passenger vehicle fleet-Analysis of trends in the technological profile and related impacts on CO2 emissions. *Energy Policy, 37*(10), 3779–3789. https://doi.org/10.1016/j.enpol.2009.07.011

Mirza, A. H. (2018). Poverty Data Model as Decision Tools in Planning Policy Development. *Scientific Journal of Informatics, 5*(1), 39. https://doi.org/10.15294/SJI.V5I1.14022

Newstead, S., & D'Elia, A. (2007). *An investigation into the relationship between vehicle colour and crash risk*. Monash University Accident Research Centre, 263, 1–20. www.monash.edu.au/muarc

Olga. (2019). *Tren Mobil Bekas di Indonesia, Primadona Karena Beragam Tujuan*. Caroline.Id. https://www.caroline.id/berita/tren-mobil-bekas-di-indonesia/

Pal, N., Arora, P., Sundararaman, D., Kohli, P., & Palakurthy, S. S. (2017). How much is my car worth? A methodology for predicting used cars prices using Random Forest. *Advances in Intelligent Systems and Computing, 886*, 413–422. https://arxiv.org/abs/1711.06970v1

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining, 1*(1), 27. https://doi.org/10.24014/ijaidm.v1i1.4903

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology, 4*(7), 753–764.

Putro, B. E., & Saepurohman, T. (2020). *A Classification Approach to Predicting Beef Knuckle Quality using the Decision Tree and Naïves Bayes Method: Case Study: Tiga Bersaudara Factory*. 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA), 779–783. https://doi.org/10.1109/ICIEA49774.2020.9102019

Ramalingam, S., & Rajendran, S. (2019). Assessment of performance, combustion, and emission behavior of novel annona biodiesel-operated diesel engine. *Advances in Eco-Fuels for a Sustainable Environment*, 391–405. https://doi.org/10.1016/b978-0-08-102728-8.00014-0

Rogan, F., Dennehy, E., Daly, H., Howley, M., & Ó Gallachóir, B. P. (2011). Impacts of an emission based private car taxation policy - first year ex-post analysis. *Transportation Research Part A: Policy and Practice, 45*(7), 583–597. https://doi.org/10.1016/j.tra.2011.03.007

Samruddhi, K., & Ashok Kumar, R. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. *International Journal of Innovative Research in Applied Sciences and Engineering, 4*(2), 629–632. https://doi.org/10.29027/ijirase.v4.i2.2020.629-632

Sekaran, U., & Bougie, R. (2016). *Research Method for Business Textbook (A Skill Building Approach)* 7th Edition. In United States: John Wiley & Sons Inc.

Sera. (2021). *The Impacts of PPnBM Relaxation on the Used Car Industry - PT. Serasi Autoraya*. https://www.sera.astra.co.id/news/2021/03/dampak-relaksasi-ppnbm-terhadap-industri-mobil-bekas

Shin, S.-Y. S. (2013). Correlation between Car Accident and Car Color for Intelligent Service. *Journal of Intelligence and Information Systems, 19*(4), 11–20. https://doi.org/10.13088/JIIS.2013.19.4.011

Suryantoro, M. T., Sugiarto, B., & Mulyadi, F. (2016). Growth and characterization of deposits in the combustion chamber of a diesel engine fueled with B50 and Indonesian biodiesel fuel (IBF). *Biofuel Research Journal, 3*(4), 521–527. https://doi.org/10.18331/BRJ2016.3.4.6

Syapsan. (2019). The effect of service quality, innovation towards competitive advantages and sustainable economic growth. *Benchmarking: An International Journal, 26*(4), 1336–1356. https://doi.org/10.1108/bij-10-2017-0280

Tamara, N. H. (2020). *Peta Baru Persaingan Bisnis Mobil di Indonesia - Analisis Data Katadata*. Katadata.Co.Id. https://katadata.co.id/zimi95/analisisdata/5e9a57afaf667/peta-baru-persaingan-bisnis-mobil-di-indonesia

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(2), e1289. https://doi.org/10.1002/WIDM.1289

Wang, X., Liu, T., Zheng, X., Peng, H., Xin, J., & Zhang, B. (2018). Short-term prediction of groundwater level using improved random forest regression with a combination of random features. *Applied Water Science, 8*(5). https://doi.org/10.1007/s13201-018-0742-6

Webb, G. I. (2010). *Encyclopedia of Machine Learning*. In Encyclopedia of Machine Learning (pp. 713–732). https://doi.org/10.1007/978-0-387-30164-8

Widodo, J., Kuesar, E. J., Verma, R., Purnama, I., & Wibowo, Y. (2021). *The "New Normal" of Indonesia Used Car Industry*. https://news.olx.co.id/proyeksi-bisnis-mobil-bekas-menghadapi-new-normal/

Zainuri, F., Sumarsono, D. A., Adhitya, M., & Siregar, R. (2017). *Design of synchromesh mechanism to optimization manual transmission's electric vehicle*. AIP Conference Proceedings, 1823. https://doi.org/10.1063/1.4978104

Zhang, H., & Li, D. (2007). *Naïve Bayes text classifier*. Proceedings - 2007 IEEE International Conference on Granular Computing, GrC 2007, 708–711. https://doi.org/10.1109/GRC.2007.4403192