p-ISSN: 2252-3006
e-ISSN: 2685-2411

Implementing kNearest Neighbor Methods to Predict Car Prices

Deshiwa Budilaksana[a1], I Made Sukarsa[a2], Anak Agung Ketut Agung Cahyawan Wiranatha[a3]
[a]Department of Information Technology, Faculty of Engineering, Udayana University
Bukit Jimbaran, Bali, Indonesia, telp. (0361) 701806
e-mail: [1]deshiwabudi@gmail.com, [2]sukarsa@unud.ac.id, [3]agung.cahyawan@unud.ac.id

Abstract

The demand for automotive in Indonesia has never subsided, considering that the human need for transportation greatly affects people's daily lives. Various attempts are made by manufacturers to produce cars of a quality that is comparable to the costs incurred and following market demand. Prediction is a process that can be done to achieve this goal. One of the prediction methods that can be used in this case is the kNearest Neighbor. The prediction process consists of a preprocessing stage that cleans and filters unnecessary variables, followed by a variable multicollinearity test stage with Variance Inflation Factor (VIF). The multicollinearity test found 4 variables that had a specific influence in predicting the VIF value of these variables, respectively 2.22, 2.08, 1.53, 1.10 for Horse Power, Car Width, Highend, and, Hatchback respectively. The four variables of the VIF test results have a positive correlation with the price variable as the dependent variable. The prediction model is made using 4 variables selected based on the VIF test, to determine the accuracy of the method used, the Linear Regression model and, the kNearest Neighbor through the validation test with Mean Absolute Error (MAE) and R^2. The kNearest Neighbor method produces an MAE test of 0.06 and R^2 results are 0.843. This can be concluded if the overall kNearest Neighbor method has qualified performance in making predictions with continuous value variables or in other words using the concept of regression.

Keywords : *prediction, productions, linear regression, knearest neighbor, multicolinearity*

1. Introduction

The business world is full of competition which requires companies to be more innovative in designing a product so that it is always the choice of consumers [1]. The demand for automotive and transportation in the world has never subsided, considering that human transportation needs are very influential in people's daily activities. This high need has an impact on the fast-growing automotive market. Product innovation that came by itself spurred the competition. Before starting to expand in the automotive business, a company needs to define how existing market conditions are to match the products to be marketed and market desires [2].

Table 1. Indonesia Car Production

Year	Production Quantity	National Sales	Export	Import
2015	1.098.780	1.103.618	207.691	82.533
2016	1.177.797	1.026.716	194.395	75.571
2017	1.216.615	1.079.534	231.169	88.683

Table 1 shows the car production in Indonesia collected from Gaikindo, from 2015 to 2017 it has always increased every year. Most production occurred in 2017, amounting to 1,216,615 units. This proves that over the years car production in Indonesia is still increasing. This causes car manufacturers to always compete to innovate in exploiting the potential of the

Indonesian market. Externally, car production is influenced by several factors, including import conditions, foreign investment and nominal wages of workers [3].

Various attempts are made by producers in balancing production costs with selling prices and on the other hand, the products produced are also by their market niche. This information can be obtained by making predictions. The production process is the spearhead of the company in making a profit. The production process is expected to produce products in accordance with company standards. One of the ways that companies can attract consumer interest is by offering quality products with appropriate service and prices. The need for good quality products makes companies design products that can compete and survive in the midst of business competition [4].

Prediction is the process of predicting something for future purposes that refers to current data to reduce uncertainty [5]. The prediction process requires data that is continuous or data that describes the measurement results so that an unlimited range of values is obtained, for example, height data, as well as discrete data or data describing categories with complete values such as blood group data. Prediction using data like this will be better using the concept of regression. Regression analysis is a method of predicting the value of a variable based on one or more independent variables. Linear regression consists of two techniques, simple and multiple. Simple regression is a regression analysis that is carried out using one independent character to estimate the value of the dependent variable. In contrast to multiple linear regression, which uses several independent variables to determine the value of one dependent variable. Industry and business have used prediction science as a tool to increase considerably in making decisions to minimize losses. Predictions can be used in various fields, one of which is for the automotive industry. An example of a suitable method for predicting is regression, where the regression method is a statistical tool to determine the effect of one or more variables called independent variables on a variable called the dependent variable [6]. kNearest Neighbor (KNN) is one of the supervised classification and prediction algorithms with a simple lazy learner concept that is easy to apply. The application of KNN can be used in a regression concept with the Unsupervised kNearest Neighbor approach, wherein regression predicts the output value. The working principle of the KNN is based on finding the closest distance between the data to be evaluated and the closest “neighbor” to the training data [7].

When an information system has a data type with a large amount of data in it, and when using the software the user needs to choose the data to be used, the prediction system becomes a solution for users to get the required data quickly [8]. Popular prediction methods is k-Nearest Neighbor [9]. The prediction process can be done using independent variables which will affect the value of the variable to be predicted. The use of data in the prediction process will affect the results given. The effectiveness of the system will be reduced if the data used is still raw and there are many disruptions to the data. Data quality is one of the determinants of the good or bad performance of a prediction system. Variation is the biggest obstacle in data usage, errors such as incomplete, inconsistent, duplication, and value conflicts will have a big influence on the prediction process. Data cleaning becomes a separate process that consists of several stages including standard determination, error detection, and error correction so that there are no anomalies that can affect the prediction results [10].

This research was conducted using car production data of 205 types of cars provided by UCI [11] to predict car prices using the KNN method, the selection of the KNN method is based on its simplicity and application of KNN's lazy learning concept in computation on continuous data such as those used for regression. Before starting the prediction process, the available data must go through preprocessing stage check to prevent bias in predictions and improving overall performance, and choose the most effective variables to be able to provide predictions. Variance Inflation Factor (VIF) is used to measure the amount of change in the independent variable that will affect the dependent variable, while the Recursive Feature Elimination (RFE) is used to sort features based on their relevance to the dataset. The selection of features will explain the multicollinearity of the data and can affect the performance of the prediction system design. The prediction testing system is carried out to measure the value error using the Mean Absolute Error (MAE) and the Determination Coefficient using R^2.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

2. Research Method

The research starts from defining the problem, wherein this case is predicting car prices. Next is collecting data and also studying literature related to prediction systems such as the kNearest Neighbor algorithm used and how to handle data and collect it.

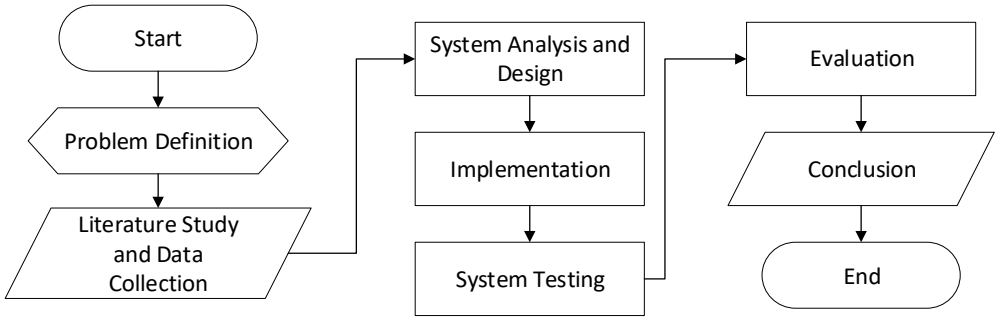


Figure 1. Research Process

Figure 1 shows the case uses secondary data on car prices where this data has a continuous variable suitable for prediction using the Regression approach. After the data and related literature have been collected, the next step or the third step is to determine how to use the kNearest Neighbor method to solve the problem of predicting car prices by utilizing existing data. The fourth is the implementation process of the previously formulated analysis and design. The fifth step is the testing phase of the system that has been implemented and then the sixth step is to evaluate the system designed where the aspects evaluated are predictive accuracy and implementation performance as well as entering the final stage or the seventh stage, namely concluding and providing suggestions for the research that has been carried out [12].

2.1. Description

The car price prediction system is made using the kNearest Neighbor method. Before being able to produce predictions, it is necessary to carry out the data training stage and the data testing phase. Both stages are carried out based on a dataset that has been divided into two for each stage, namely training data and test data. Previously, this data will enter the preprocessing stage where one of the goals is to clean the data from noise or interference when making predictions and reduce the accuracy of the results.

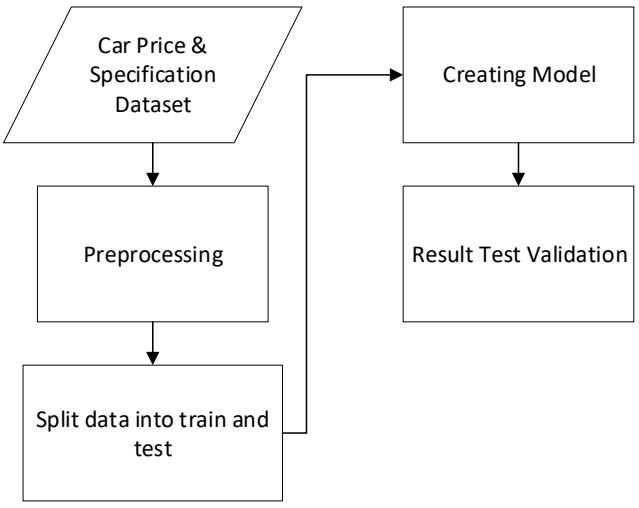


Figure 2. Description

p-ISSN: 2252-3006
e-ISSN: 2685-2411

In general, the system requires input to make predictions. This case uses a dataset of car prices provided by UCI. The total data contained in this dataset is 205 rows of data.

2.2. Data Cleansing

Data Cleansing is the process of cleaning data which includes the detection and repair of datasets, tables, and databases that are inaccurate, corrupt, or incomplete or can be called dirty data. The “dirty” data will later be replaced, modified, or deleted so that the available data is not ambiguous or conflict occurs when it is being processed. One example of a common conflict when data is still dirty is the appearance of the same name (Homonym) on 2 different objects and different names for the same object which makes the data overlap.

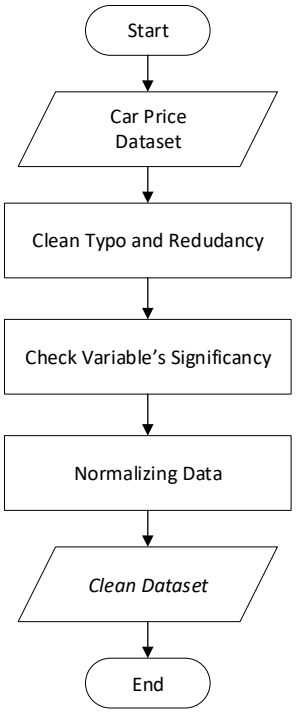


Figure 3. Data Cleansing

The Data Cleansing process used is shown in Figure 3, where the initial dataset which was still "dirty" cleaned, starting from checking for typos or typing errors and data redundancies, then sorting out variables that had a significant effect on price changes through visualization. Finally, the data that has been processed is stored as new, “clean” data.

2.3. Prediction

Activities that aim to predict something that will happen in the future are a common understanding of prediction . The use of predictions in the economic sector is to calculate the prospects of a business. Prediction calculations are based on data collected over a certain period of time [13]. Predictions can be qualitative or quantitative in nature, but qualitative calculations have drawbacks in calculating certainty because the variables are very relative in nature, on the other hand, quantitative predictions depend heavily on the use of variables and methods to get the best results [14]. Designing a prediction method must go through three stages which are as follows [15].

1. Analysis of past data to get an overview of the data patterns in question.
2. Choose the correct prediction method. Each prediction method has its own capabilities and of course, the right method will produce the best forecast, to test forecasting is to calculate the error value. The smaller the error value generated, it can be assumed that

the prediction is more accurate, but 100% accuracy indicates overfitting of the prediction model.

3. Transform the data according to prediction needs and adjust it according to the needs of the method. If necessary, data changes can be made as needed.

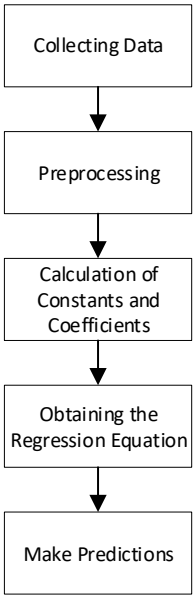


Figure 4. Prediction Process

The prediction process begins with collecting data. The data collected is car production data from various automotive brands. The data obtained then enters the preprocessing stage where at this stage the data will be cleaned, including filtering, which is removing unnecessary variables and classifying car types into low, medium, and high classes, also cleaning errors in writing brand names. Besides that, preprocessing is also useful for cleaning null data. After going through the preprocessing, the data is then divided into test data and training data before the modeling process is carried out using training data. After the model is found, the next process is to test the model using test data. Aspects of the prediction system that will be measured are accuracy using Mean Absolute Error, and R^2 Score and performance against time.

3. Literature Review

The literature review contains references to supporting literatures that are used in general in this research process. Libraries obtained from various reference sources contained in this study include marketing, data visualization, data preprocessing, data mining, linear regression, mean absolute error, mean squared error and R^2 score.

3.1. Production

Production is an activity to create, create and produce. Production can only be done if the raw materials for production are fulfilled. The resources needed to be able to carry out production consist of human resources, natural resources, capital resources in all their forms, and the skills of all these resources which are also called production factors. Production has a function of the maximum amount of output produced with a particular input combination. This function shows the nature of the relationship between the factors of production and the level of products produced. There are several variables that can affect production in industry, including working capital that can support the operations of a business, raw materials which are the materials that make up a large part of the finished product or you could say the main ingredient of a product, labor which is a term for those who work in a company and is tasked with processing raw materials into finished products and the fourth variable is a market where the

interaction between demand and supply occurs so that the balance of price and quantity is the key in pricing [16].

3.2. Multicollinearity

Multicollinearity is one of the classic assumptions in linear regression analysis where there is a tendency for a relationship between two or more independent variables. The purpose of conducting this test is to select the variables that will be the predictors. The high correlation between variables causes an unclear relationship between the dependent variable and the independent variable on the model value. When two or more independent variables have an attachment, the contribution of the variable in providing predictions will multiply and cause bias. Multicollinearity prevention can be done by calculating the Variance Inflation Factor (VIF). The way the VIF works is to measure how much the variance of the independent variable increases in a linear fashion compared to other independent variables. Multicollinearity can consist of several conditions, VIF> 10 indicates strong multicollinearity, VIF> 5 indicates moderate collinear predictor variables with other independent variables, and if the value of VIF> 1 indicates small collinearity. One of the decisions that can be taken if there is multicollinearity is to eliminate variables according to the VIF provisions to prevent bias in the designed model [17].

3.3. *Data Preprocessing*

Data Preprocessing is the process of preparing raw data before the transformation process is carried out into a form that has been adjusted to the needs of data users Data preprocessing is carried out considering the data obtained from the database tends to be raw. While the prediction algorithm works with data that is formatted in a certain way before the training data process begins, this uniformity is intended to make it easier for data to be read and progressively. Data non-uniformity can be classified into 3, namely incomplete data, noisy data, and inconsistent data. Therefore, before the next stage is carried out, it is necessary to convert the data into an appropriate format. [18].

3.4. kNearest Neighbor

kNearest Neighbor (KNN) is a supervised method (requires training data) that is quite simple to apply. KNN is known as the laziest learning method wherein studying a data it is not by forming a certain model or function, but rather provides the closest k value from the training data that has the best similarities [19]. After getting the k value, the most votes achieved will be entered directly into the test data [20]. The popularity of the KNN is comparable to its simplicity of storing multiple sample data collections where k number of available samples can also be referred to as the nearest neighbor based on distance measures such as the Euclidean Distance and make predictions based on it [21].

Regression prediction with KNN is obtained based on the total sum of all k neighbors where this weight is inversely proportional to the distance from the Euclidean Input. The Euclidean Distance function can be seen in the following function.

$$E(x,p) = \sqrt{\sum_{a}^{m}(x_a - p_a)^2} \quad (1)$$

Based on Formula 1, it can be explained if x and p are query points or the distance to be sought from one point to another. This method is sensitive to the number of "neighbors" selected and this is the uniqueness of this method [22].

3.5. Mean Absolute Error

Steps were taken to determine the accuracy of the recommendation results using the Mean Absolute Error (MAE), where MAE calculates the absolute value of the average difference obtained from the predicted value compared to the actual value [23]. Accuracy Test with Mean Absolute Error has the following equation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|fi - yi| \quad (2)$$

Information:

MAE = Mean Absolute Error
fi = Actual Value
yi = Predicted Value

3.6. Coefficent of Determination (R^2)

Analysis of R Square (R^2) or it can be called the coefficient of determination is used to determine the percentage of the relationship between the independent variable and the dependent variable simultaneously, or in other words, the coefficient of determination can determine how much influence the independent variables give together in determining the value. dependent variable. The coefficient values range from 0 to 1. If the R^2 value is close to 1 and not less than 0.5, it can be said that the influence of the independent variable is strong to affect the value of the dependent variable. This indicates that the prediction system can explain well the variables used in its duration. The types of coefficient of determination relationship can be seen in Table 1 Analysis of R Square (R^2) or it can be called the coefficient of determination is used to determine the percentage of the relationship between the independent variable and the dependent variable simultaneously, or in other words, the coefficient of determination can determine how much influence the independent variables give together in determining the value. dependent variable. The coefficient values range from 0 to 1. If the R^2 value is close to 1 and not less than 0.5, it can be said that the influence of the independent variable is strong to affect the value of the dependent variable. This indicates that the prediction system can explain well the variables used in its duration. The types of coefficient of determination relationship can be seen in Table 2 [24].

Table 2. Coefficient Interpretation

Coefficient Interval	Relation
0,8 – 1	Very Strong
0,6 – 0,79	Strong
0,4 – 0,59	Strong Enough
0,2 – 0,39	Weak
0 – 0,19	Very Weak

$$R^2 = 1 - \frac{(n-k-1)S^2_{y.12\ldots k}}{(n-1)s^2_y} \quad (3)$$

Formula 3 shows the method used to obtain the coefficient of determination, with the following information.
R^2 = coefficient of determination
Sy = Standard definition of the dependent variable Y
n = Number of samples

4. Result and Discussion

Results and discussion explain the results of testing car price predictions using the kNearest Neighbor method.

4.1. Data Preprocessing

As previously explained, the data used in this study is a dataset of 205 cars with the categories they own. The following is a list of categories in the dataset.

Table 3. Dataset's Variables

Variable's Name			
Symboling	Drive Wheel	Engine Type	Horse Power
Fuel Type	Wheel Base	Cylinder Number	City MPG
Aspiration	Car Length	Engine Size	Highway MPG
Car Body	Car Width	Bore Ratio	Curb Weight
			Price

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Preprocessing is done to remove problems with data that can interfere with the next process. In total 16 variables can be used to determine the price variable or car price as shown in Table 3. Preprocessing will be carried out to check and at the same time improve if there is incomplete data either by combining data, manipulating data, or even deleting data if necessary. The first process that will be passed is cleaning the data from writing errors or typos. The following is the process of clearing the typo in the dataset.

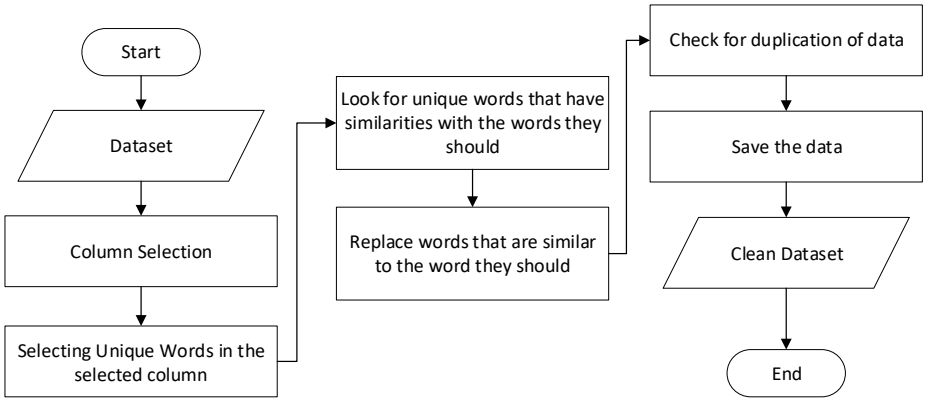


Figure 5. *Cleaning Typos*

Cleaning writing errors starts from selecting the column to be cleaned as shown on Figure 5, in this case, the column to be cleaned is the name of the car brand. Next is to select the unique words in the column, here you will see what brands are included in the dataset and the wrong writing can also be seen from the similarities between words. The examples of writing error conditions can be seen in the image below.

```
9           audi
200        volvo
160       toyota
170       toyota
119     plymouth
139       subaru
79     mitsubishi
182     vokswagen
54          mazda
188    volkswagen
```

Figure 6. Writing Error

Figure 6 shows that there is an error in writing the brand name in rows 182 where the brand name is written as 'vokswagen' while the actual brand name is 'volkswagen' as written in rows 188. The first step to clear the typo is to check the brand name column with the data. unique, unique column selection results can be seen in Table 4.

Table 4. Unique Car Brand

Unique Car Brand		
alfa-romero	audi	bmw
chevrolet	dodge	honda
isuzu	jaguar	mazda
buick	mercury	mitsubishi
nissan	peugeot	plymouth

p-ISSN: 2252-3006
e-ISSN: 2685-2411

porsche	renault	saab
subaru	toyota	volkswagen
volvo	maxda	porcshce
toyouta	vokswagen	vw

Based on table 4 it can be seen if there is a name writing error in the names of several brands such as 'maxda' for 'mazda', 'vokswagen' and 'vw' for 'volkswagen' and 'porcshce' for 'porsche'. Displaying all of the unique brand name data will help make it easier to identify typographical errors that have occurred. After sorting out what words are typos, the next process is to replace those words with the proper brand name words. Before the clean data is stored, check whether there is any duplication in the brand name or not. Table 5 shows the clean table of brand names.

Table 5. Clean Car Brand

Car Brand Name		
alfa-romero	audi	bmw
chevrolet	dodge	honda
isuzu	jaguar	mazda
buick	mercury	mitsubishi
nissan	peugeot	plymouth
porsche	renault	saab
subaru	toyota	volkswagen
volvo		

After the writing errors are corrected, the next process is to see whether there are variables that are interrelated and can form a new variable or it can be called a correlation. Correlation is the relationship between two variables influencing each other. Correlation consists of Positive Correlation, Negative Correlation, and No Correlation. This research will show how each variable (correlation) affects the rise or fall of car prices. First, based on the dataset, the distribution of car prices is as follows. The key to this process is data visualization. The City MPG and Highway MPG variables have a similar pattern and also have the same unit of measurement, the correlation value of the two variables shows the number 0.841 which indicates that the two variables have a positive correlation which is evident from the similarity of the pattern which can be seen in Figure 6.

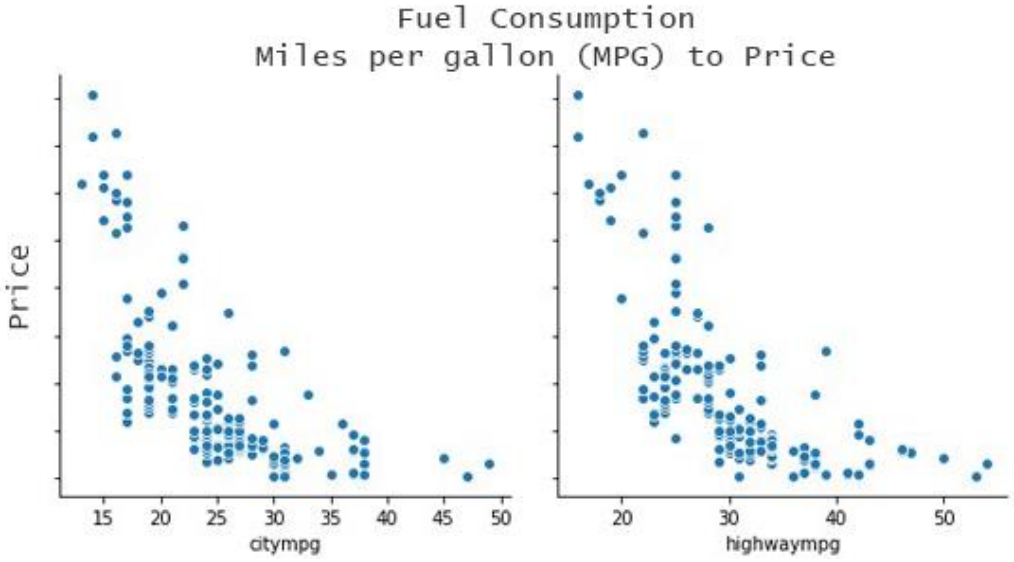
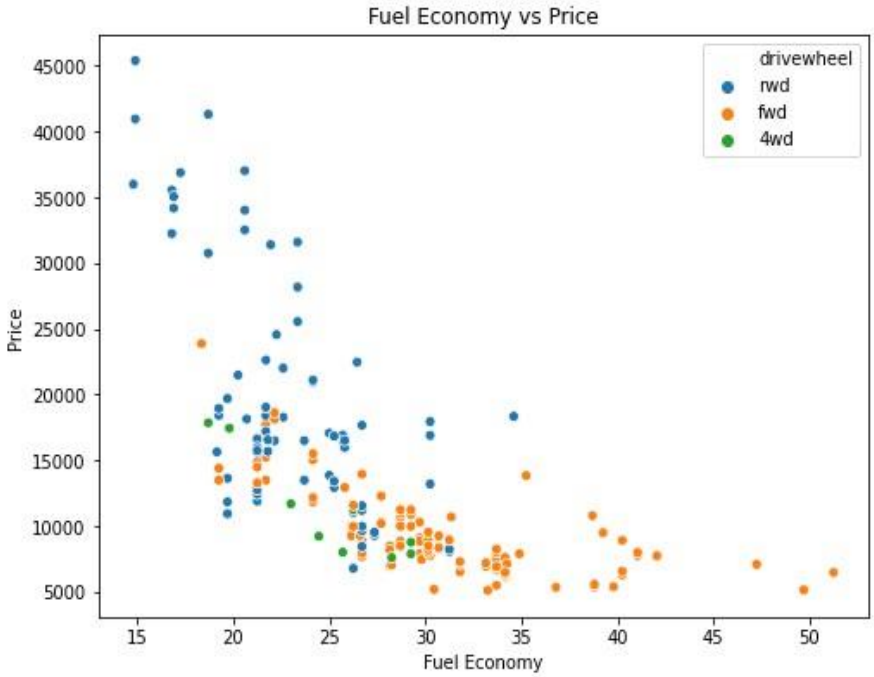


Figure 7. *City & Highway MPG*

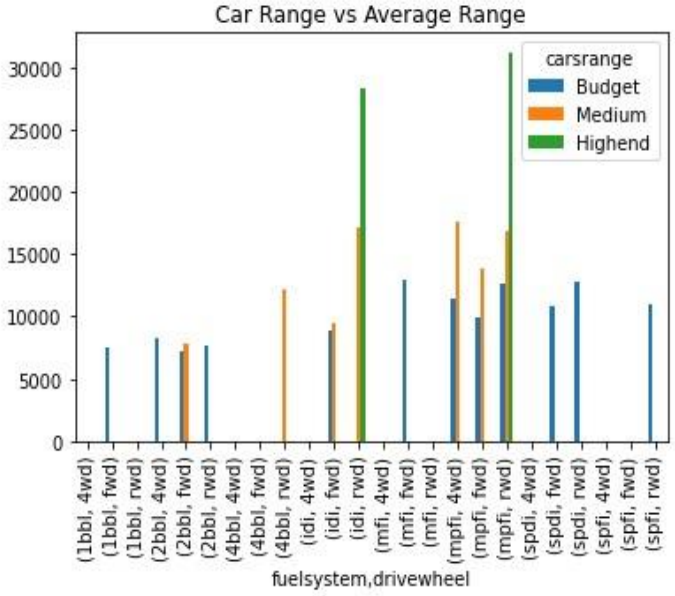
p-ISSN: 2252-3006
e-ISSN: 2685-2411

These two variables show the amount of fuel the vehicle consumes in Gallons against the Mile traveled (Miles per Gallon). Based on these facts, the combination of the two variables will be called Fuel Economy to show the fuel efficiency of the car. Fuel Economy data can be seen in Figure 7.



Gambar 8. *Fuel Economy*

Based on the variable Price or car price, several cars can be classified into 3 groups, namely 'Budget', 'Medium' and 'Highend'. Classification of cars based on Fuel Systems and Wheel drives can be seen in Figure 8.



Gambar 9. *Car Range*

After going through this preprocessing stage, it can be seen that there are 2 new variables in the dataset, namely Fuel Economy and Car Range. After adding variables, the next

p-ISSN: 2252-3006
e-ISSN: 2685-2411

step is to determine what variables can be used and what variables will actually interfere with the prediction results using Recursive Feature Elimination (RFE). Feature selection refers to its relevance in the dataset. The number of features can affect how long the prediction process runs, so the features used must also be selected efficiently so that predictions are not only accurate but also have faster performance. Based on the total 18 variables currently present in the dataset, the results of selecting features using RFE are as follows.

Tabel 6. RFE Variables

Variables Name		
Curb Weight	Hatch Back	Twelve
Horse Power	Sedan	Highend
Fuel Economy	Wagon	
Car Width	DOHCV	

Table 5 shows the variables resulting from the feature elimination process with RFE. After feature elimination was carried out with RFE, the next process was to check the multicollinearity of each variable.

4.2. Variables Multicollinearity

Multicollinearity is one of the assumptions in Linear Regression where there is a relationship between independent variables that influence each other's values. One method that can be used to check multicollinearity is the Variance Inflation Factor (VIF). The way VIF works is by comparing the variance in the independent variable. The following are the results of the VIF test to check the multicollinearity of variables in the dataset with variables that have been selected by RFE outside of the constant value. The following is the VIF value before testing.

Tabel 7. Variable's VIF

Nb	Variable	VIF	Nb	Variable	VIF
1	Curb Weight	8.33	5	Horse Power	5.06
2	Sedan	6.13	6	Wagon	3.58
3	Hatchback	5.67	7	Fuel Economy	3.56
4	Car Width	5.19	8	Highend	1.68
			9	DOHCV	1.62

The VIF value in Table 7 shows several values that are more than 5 such as the Horse Power, Seda, and Curb Weight variables with the respective values being 5.06, 5.19, 5.67, 6.13, and 8.33.

Table 8. Variance Inflation Factor Test

No	Variabel	VIF
1	Horse Power	2.22
2	Car Width	2.08
3	Highend	1.53
4	Hatch Back	1.10

Table 8 shows VIF after eliminating variables that have multicollinearity. Seeing the correlation of each variable to car prices as the dependent variable, Horse Power tends to increase car prices along with the increasing value of HorsePower. The Car Width variable also affects the car price value positively, namely that the price value will increase as the Car Width value increases. High End and Hatch Back variables are dummies or categorical variables. Examples of clean data are as follows.

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Table 9. Clean Data

	Horse Power	Car Width	Hatchback	Highend
1	0.116129	0.2	0	0
2	0.212903	0.315789	0	0
3	0.206452	0.421053	1	0
4	0.387097	0.157895	0	0
5	0.135484	0.136842	1	0

Apart from eliminating multicollinearity interference, VIF is also used to select what features will be used in the model so that it is not only able to eliminate interference in the prediction model, but also can help speed up the prediction performance.

4.3 Validation Test

Prediction model testing is done by using the Mean Absolute Error and the Coefficient of Determination (R^2). Testing the Mean Absolute Error and R^2 on the KNN method obtained results, namely 0.6 and 0.843 for MAE and R^2 respectively, this indicates that the independent variables used to make predictions have a very strong attachment in influencing the resulting value in the independent variable.

5. Conclusion

Based on the discussion regarding the implementation of the KNN method to predict car prices, the conclusions obtained are as follows. Before the prediction is obtained, it is necessary to carry out the preprocessing stage. This preprocessing stage corrects blank data as well as writing errors. Preprocessing also checks the correlation between variables, at this stage, it produces two combined variables, namely Fuel Economy which is generated from the City MPG correlation, and Highway MPG which shows a correlation value of 0.841. Another variable that is produced is a car class classification called Cars Range which classifies cars into Budget, Medium, and Highend. Recursive Feature Elimination (RFE) is used to eliminate unused independent variables. Elimination of variables is done to determine which variables have a significant influence in predicting and which variables are less significant and can even interfere with the prediction process. Before looking at the RFE value, a dummy variable is created to create categorical variables. There are 10 total variables included in the significant variables according to the RFE test which can be seen in Table 2.

After the preprocessing stage, the next step is to check the Multicollinearity of the Variable. Multicollinearity can bias predictions. Multicollinearity was tested using the Variance Inflation Factor (VIF). The results of the VIF test resulted in 4 variables with a VIF value of more than 5, namely Horse Power, Car Width, High End, and Hatch Back with values respectively 2.22, 2.08, 1.53, and 1.10. These four variables are used in the prediction model. Model validation is done by testing Mean Absolute Error and R^2. The result of the Mean Absolute Error test is 0.6 and the R^2 value is 0.843. This indicates that the kNearest Neighbor method has tolerable accuracy to be used in providing predictions on data with continuous variables as is commonly used in the regression method.

References

[1] Murni Marbun, “Implementasi Sistem Informasi Penjualan Mobil Dengan Metode Feature Driven Development (Fdd) Pada Pt.Capella Medan,” *Jurnal Mantik Penusa*, vol. 2, no. 1, hal. 15–21, 2015.

[2] S. N. Untari, S. Djaja, dan J. Widodo, “Strategi Pemasaran Mobil Merek Daihatsu Pada Dealer Daihatsu Jember,” *JURNAL PENDIDIKAN EKONOMI: Jurnal Ilmiah Ilmu Pendidikan, Ilmu Ekonomi dan Ilmu Sosial*, vol. 11, no. 2, hal. 82, 2018, doi: 10.19184/jpe.v11i2.6451.

[3] M. Hanif dan S. R. T. Astuti, “ANALISIS PENGARUH MOTIVASI KONSUMEN, PERSEPSI KUALITAS, SIKAP KONSUMAN DAN CITRA MEREK TERHADAP KEPUTUSAN PEMBELIAN DENGAN MINAT BELI SEBAGAI VARIABEL INTERVENING (Studi Pada Calon Konsumen Mobil Datsun di Kota Semarang),”

DIPONEGORO JOURNAL OF MANAGEMENT, vol. 7, no. 4, hal. 1–12, 2018.

[4] L. Utari dan N. Triyanto, "Prediksi Jumlah Produksi Mobil Pada Perusahaan Karoseri Dengan Menggunakan Metode Exponential Smoothing," *Teknois : Jurnal Ilmiah Teknologi Informasi dan Sains*, vol. 7, no. 1, hal. 59–67, 2019, doi: 10.36350/jbs.v7i1.34.

[5] P. Katemba dan R. K. Djoh, "Prediksi Tingkat Produksi Kopi Menggunakan Regresi Linear," *Jurnal Ilmiah FLASH*, vol. 3, no. 1, hal. 42–51, 2017.

[6] R. Ishak, "Prediksi Jumlah Mahasiswa Registrasi Per Semester," *ILKOM Jurnal Ilmiah*, vol. 10, no. 2, hal. 136–143, 2018.

[7] M. Nanja dan P. Purwanto, "Metode K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Harga Komoditi Lada," *Pseudocode*, vol. 2, no. 1, hal. 53–64, 2015, doi: 10.33369/pseudocode.2.1.53-64.

[8] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, dan L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Computer Science*, vol. 1, no. 2, hal. 2811–2819, 2010, doi: 10.1016/j.procs.2010.08.006.

[9] S. Wiyono dan T. Abidin, "Implementation of K-Nearest Neighbour (Knn) Algorithm To Predict Student'S Performance," *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 9, no. 2, hal. 873–878, 2018, doi: 10.24176/simet.v9i2.2424.

[10] N. Putu, A. Widiari, I. M. Agus, D. Suarjaya, dan D. P. Githa, "Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali," vol. 8, no. 2, hal. 137–145, 2020.

[11] D. Dua dan C. Graff, "UCI Machine Learning Repository," *University of California, Irvine, School of Information and Computer Sciences*, 2017. [Daring]. Tersedia pada: http://archive.ics.uci.edu/ml/datasets/Automobile.

[12] K. S. A. Wasika, "Klasifikasi Kunci Gitar Menggunakan Spectral Analysis," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, vol. 8, no. 1, hal. 61–71, 2020.

[13] M. Hakimah, R. R. Muhima, dan A. Yustina, "Rancang Bangun Aplikasi Persediaan Barang Dengan Metode Trend Projection," *SimanteC*, vol. 5, no. 1, hal. 37–48, 2015.

[14] G. N. Ayuni dan D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," vol. 14, no. 2, hal. 79–86, 2019.

[15] T. Indarwati, T. Irawati, dan E. Rimawati, "Penggunaan Metode Linear Regression Untuk Prediksi Penjualan Smartphone," *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 6, no. 2, hal. 2–7, 2019, doi: 10.30646/tikomsin.v6i2.369.

[16] Z. V. Sumolang, T. O. Rotinsulu, dan D. S. M. Engka, "Analisis Faktor-Faktor Yang Mempengaruhi Produksi Industri Kecil Olahan Ikan Di Kota Manado," *Jurnal Pembangunan Ekonomi Dan Keuangan Daerah*, vol. 19, no. 3, hal. 1–17, 2019, doi: 10.35794/jpekd.16459.19.3.2017.

[17] E. Supriyadi, S. Mariani, dan Sugiman, "PERBANDINGAN METODE PARTIAL LEAST SQUARE (PLS) DAN PRINCIPAL COMPONENT REGRESSION (PCR) UNTUK MENGATASI MULTIKOLINEARITAS PADA MODEL REGRESI LINEAR BERGANDA," *UNNES Journal of Mathematics*, vol. 6, no. 2, hal. 117–128, 2017.

[18] I. P. Arya, P. Wibawa, I. K. A. Purnawan, D. Purnami, S. Putri, dan N. Kadek, "Prediksi Partisipasi Pemilih dalam Pemilu Presiden 2014 dengan Metode Support Vector Machine," *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, vol. 7, no. 3, hal. 182–192, 2019.

[19] F. Astutik, "Sistem Pengenalan Kualitas Ikan Gurame Dengan Wavelet, Pca, Histogram Hsv Dan Knn," *Lontar Komputer*, vol. 4, no. 3, hal. 336–346, 2015, doi: 10.24843/LKJITI.

[20] S. P. Poornima, C. N. Priyanka, P. Reshma, S. K. Jaiswal, dan K. N. Surendra Babu, "Stock price prediction using KNN and linear regression," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 5 SpecialIssue, hal. 142–145, 2019.

[21] F. Martínez, M. P. Frías, F. Charte, dan A. J. Rivera, “Time Series Forecasting with KNN in R : the tsfknn Package,” vol. 11, no. December, hal. 229–242, 2019.

[22] N. L. W. S. R. Ginantra, “Deteksi Batik Parang Menggunakan Fitur Co-Occurence Matrix Dan Geometric Moment Invariant Dengan Klasifikasi KNN,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 7, no. 1, hal. 40, 2016, doi: 10.24843/lkjiti.2016.v07.i01.p05.

[23] L. Dzumiroh dan R. Saptono, “Penerapan Metode Collaborative Filtering Menggunakan Rating Implisit pada Sistem Perekomendasi Pemilihan Film di Rental VCD,” *Jurnal Teknologi & Informasi ITSmart*, vol. 1, no. 2, hal. 54, 2016, doi: 10.20961/its.v1i2.590.

[24] H. Pham, “A new criterion for model selection,” *Mathematics*, vol. 7, no. 12, hal. 1–12, 2019, doi: 10.3390/MATH7121215.