JITEEHA: Journal of Information Technology Application in Education, Economy, Health and Agriculture
Vol. 02, No. 01, February 2025, pp. 16~23, e-ISSN: 3090-3939
https://journal.iteeacademy.org/index.php/jiteeha/

Classification of Used Car Prices Using the Naive Bayes Method

Bintang Abillah[1], Djourdi Amrida Pratama[2], Rizandi Agung Baskara[3], Mugi Praseptiawan[4], Pauline Hanfiro[5]
[1,2,3] Department of Informatics Engineering, University of Widyagama Malang, Borobudur Street No. 35, Malang City, East Java, Indonesia
[4] Department of Informatics Engineering, Institut Teknologi Sumatera, Sumatera, Indonesia
[5]University of the South Pacific -Fiji

Article Info

Article history:

Received January 01, 2025
Revised February 10, 2025
Accepted February 30, 2025

Keywords:

Naive Bayes
Purchasing Decision Prediction
Data Mining
Used Motorcycles
Gaussian Naive Bayes
Data Analysis

ABSTRACT

This research uses the Naive Bayes algorithm to predict used car purchasing decisions based on attributes such as brand, year of production, mileage, engine condition, completeness of features, and maintenance history. By applying the Gaussian Naive Bayes approach to handling continuous data, this research aims to develop a reliable prediction model while identifying the attributes that most influence purchasing decisions. The test results show that the prediction model achieved a correct accuracy level of 80%, and an incorrect accuracy of 20%, which indicates the ability of the Naive Bayes algorithm to handle data classification. This research provides insights that can support industry players in designing more effective sales strategies based on accurate data analysis.

Corresponding Author:

Bintang Abillah
Department of Informatics Engineering
Universitas Widya Gama Malang
Jl. Borobudur No. 35 Malang – Jawa Timur, Indonesia
Email: bintangabillah01@gmail.com

1. INTRODUCTION

The rapid development of the automotive industry has encouraged an increase in the used car market in various countries. Used car buying and selling transactions reach 3-4 times compared to new cars [1]. This shows the high public interest in used cars as a more affordable alternative. However, determining the price and suitability of used engine cars is a major challenge for consumers and sellers [2]. The complexity of determining the value of a used car is influenced by various factors such as brand, year of production, kilometers traveled, engine condition, complete features, and maintenance history [3]. This uncertainty and subjectivity in assessment often causes significant price gaps in the market. The use of conventional methods for analyzing used car prices is considered less effective because it is unable to process historical data optimally [4].

As machine learning technology develops, the Naïve Bayes method has emerged as a promising approach in predictive analysis of used cars. Naïve Bayes is a probabilistic classification algorithm that is able to process multiple variables efficiently with a good level of accuracy [5]. This shows that the Naïve Bayes method is able to provide up to 85% accuracy in predicting used car prices based on various parameters [6]. The advantage of the Naïve Bayes method lies in its ability to handle large and complex datasets with relatively simple computations [7]. This algorithm uses the principle of conditional probability which allows analysis of relationships between variables independently. This proves that Naïve Bayes can identify hidden patterns in historical used car sales data detected using traditional methods [8].

The implementation of Naïve Bayes in used car analysis has shown promising results in various studies. Using Naïve Bayes to classify the condition of used cars based on 15 technical parameters with a precision level of 82% [9]. By developing a used car recommendation system based on Naïve Bayes which is able to provide purchasing suggestions according to consumer preferences and budget [10]. The development of analysis in the used car industry continues to experience significant progress. The integration of machine learning has changed

the way industry players analyze and determine the selling value of used cars [11]. The importance of a data-based approach in improving assessment accuracy and transaction efficiency in the used car market [12].

Based on the problems that have been described, this research has several main objectives. First, develop an accurate used car sales value classification system using the Naïve Bayes method. Second, identify key factors that influence the selling value of used cars through probabilistic analysis. Third, evaluate the level of accuracy and performance of the Naïve Bayes method in the context of used car appraisal. To achieve this goal, this research applies several systematic methodological stages. The first stage is collecting a dataset of used cars which includes various parameters such as brand, type, year of production, kilometers traveled and selling price. Next, data preprocessing is carried out including cleaning, normalization and data transformation to ensure optimal input quality. The implementation of the Naïve Bayes algorithm is carried out by dividing the dataset into training and testing data with a ratio of 80:20. The training process will produce a probabilistic model which is then validated using testing data. Model performance evaluation is carried out using metrics such as accuracy, precision, recall, and F1-score to measure classification effectiveness. The results of this evaluation will show how accurate the model is in predicting the selling value of used cars based on predetermined parameters.

1. RESEARCH METHODOLOGY

2. 1 Research Framework

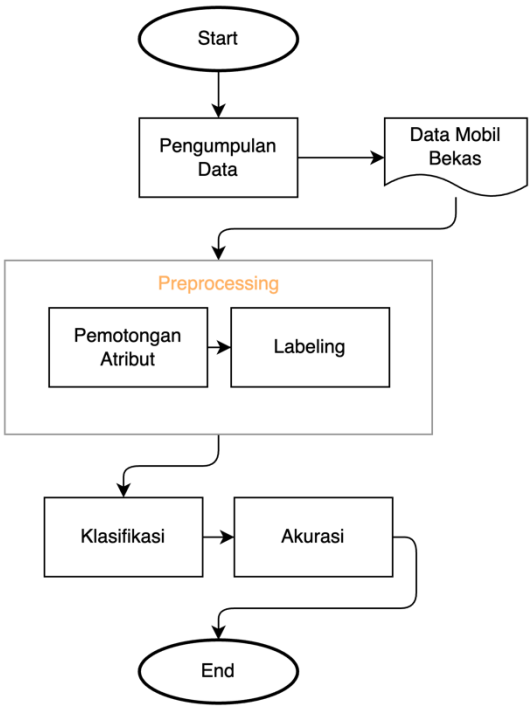

Figure 1 Framework

2. 2 Data Collection

The data used in this research was obtained through data collection from the Kaggle site, a platform that provides various quality data collections for analysis and machine learning model development. The selection of data from Kaggle was based on the relevance and completeness of the information in accordance with the objectives of this research. From the Kaggle site, a dataset of 10 used cars was obtained, with 9 attributes including model, year of production, price, transmission, mileage, fuel, tax, MPG and engine size.

2. 3 Preprocessing

The data that has been collected will then go through a series of preprocessing processes which aim to clean and present the data in optimal conditions for the classification process. The first stage in this process is to cut or delete several attributes that are not very relevant for calculations. Previously the dataset had 9 attributes, but in this process 5 attributes will be removed in order to maximize the classification process. Removed attributes include transmission, mileage, fuel, tax, and MPG.

The second stage in the preprocessing process is labeling the dataset. Each data is given a label Cheap or Expensive based on the attributes it has. This process functions for the model training process. The following is the form of the dataset that has gone through preprocessing:

Table 1. Data After Preprocessing

Car	Used Car Value (*Training Data*) Attribute				Class
	Model (C1)	**Production Year(C2)**	**Price(C3)**	**Engine Capacity (C4)**	**Value (C5)**
1	Toyota Kijang Innova	2007	Rp80,000,000.00	2.5	Cheap
2	Toyota Avanza	2012	Rp105,000,000.00	1.3	Cheap
3	Toyota Avanza	2012	Rp105,000,000.00	1.3	Cheap
4	Toyota Fortuner	2012	Rp220,000,000.00	2.5	Cheap
5	Toyota Avanza	2012	Rp157,000,000.00	1.3	Cheap
6	Toyota Kijang Innova	2016	Rp177,000,000.00	2.5	Cheap
7	Toyota Avanza	2016	Rp214,000,000.00	1.3	Expensive
8	Toyota Agya	2016	Rp200,000,000.00	1.3	Expensive
9	Toyota Agya	2020	Rp136,000,000.00	1.3	Expensive
10	Toyota Avanza	2020	Rp152,000,000.00	1.3	Expensive

2. RESULTS AND DISCUSSION

3. 1 NBC Clasiffication

NBC is used to predict classes from data in the testing set group. The classification process using NBC begins by determining the probability of the data for each attribute. Here's the calculation:

Table 2. Probability C1

Model	Probability C1			
	Number of Event Selected		Probabiliy	
	Yes	**No**	**Cheap**	**Expensive**
Toyota Kijang Innova	2	0	2	0
Toyota Avanza	3	2	3	2
Toyota Fortuner	1	0	1	0
Toyota Agya	0	2	0	2
Total	6	4	1	1

Table 3. Probability C2

Production Year	Probability C2			
	Number of Event Selected		Probability	
	Yes	**No**	**Cheap**	**Expensive**

2007	1	0	1	0
2012	4	0	4	0
2016	1	2	1	2
2020	0	2	0	1
Jumlah	6	4	1	1

Table 4. Probability C3

n	**Probability C3**	
	Cheap	**Expensive**
1	Rp80.000.000,00	Rp214.000.000,00
2	Rp105.000.000,00	Rp200.000.000,00
3	Rp105.000.000,00	Rp136.000.000,00
4	Rp220.000.000,00	Rp152.000.000,00
5	Rp157.000.000,00	-
6	Rp177.000.000,00	-
Average	Rp140.666.667	Rp175.500.000
Standard Deviation	Rp53.113.714	Rp37.394.295

Table 5. Probability C4

Engine Capacity	**Probability C4**			
	Number of Event Selected		**Probability**	
	Yes	**No**	**Cheap**	**Expensive**
1.3	3	4	3	4
2.5	3	0	3	0
Total	6	4	1	1

After calculating the probability, the data classification process is then carried out using NBC. Classification is carried out using Microsoft Excel with the following calculations:

Table 6. Testing Data

Car	**Data Testing**				
	Attribute				**Class**
	Model (C1)	**Prduction Year (C2)**	**Price (C3)**	**Engine Capacity (C4)**	**Value (C5)**
1	Toyota Kijang Innova	2007	Rp80.000.000,00	2,5	?
2	Toyota Avanza	2012	Rp105.000.000,00	1,3	?
3	Toyota Avanza	2012	Rp105.000.000,00	1,3	?
4	Toyota Fortuner	2012	Rp220.000.000,00	2,5	?
5	Toyota Avanza	2012	Rp157.000.000,00	1,3	?
6	Toyota Kijang Innova	2016	Rp177.000.000,00	2,5	?
7	Toyota Avanza	2016	Rp214.000.000,00	1,3	?
8	Toyota Agya	2016	Rp200.000.000,00	1,3	?
9	Toyota Agya	2020	Rp136.000.000,00	1,3	?
10	Toyota Avanza	2020	Rp152.000.000,00	1,3	?

Table 7. Testing Data Classification

Testing Data Classification							
Car	**Gaussian Density Classification (C3)**		**Probability**		**Classification of Probability Values**		**Normalization Results**
	Cheap	**Expensive**	**Cheap**	**Expensive**	**Cheap**	**Expensive**	
1	0,000000003912	0,000000000409	0,000000000158	0,000000000000	1	0	Cheap
2	0,000000005995	0,000000001804	0,000000000242	0,000000000059	0,80	0,20	Cheap
3	0,000000005995	0,000000001804	0,000000000030	0,000000000013	0,70	0,30	Cheap
4	0,000000002462	0,000000005255	0,000000000025	0,000000000038	0,39	0,61	Expensive
5	0,000000007164	0,000000009440	0,000000000145	0,000000000000	1	0	Cheap
6	0,000000005944	0,000000010660	0,000000000030	0,000000000078	0,28	0,72	Expensive
7	0,000000002896	0,000000006280	0,000000000117	0,000000000411	0,22	0,78	Expensive
8	0,000000004025	0,000000008608	0,000000000020	0,000000000063	0,25	0,75	Expensive
9	0,000000007482	0,000000006107	0,000000000038	0,000000000044	0,46	0,54	Expensive
10	0,000000007342	0,000000008757	0,000000000037	0,000000000064	0,37	0,63	Expensive

3. 2 Testing Data Result

From the calculations that have been carried out, the following are the results of the data calculations:

Table 8. Testing Data Result

Testing Data Result					
Car	**Attribute**				**Class**
	Model (C1)	**Production Year (C2)**	**Price (C3)**	**Engine Capacity (C4)**	**Value (C5)**
1	Toyota Kijang Innova	2007	Rp80.000.000,00	2,5	Cheap
2	Toyota Avanza	2012	Rp105.000.000,00	1,3	Cheap
3	Toyota Avanza	2012	Rp105.000.000,00	1,3	Cheap
4	Toyota Fortuner	2012	Rp220.000.000,00	2,5	Expensive
5	Toyota Avanza	2012	Rp157.000.000,00	1,3	Cheap
6	Toyota Kijang Innova	2016	Rp177.000.000,00	2,5	Expensive
7	Toyota Avanza	2016	Rp214.000.000,00	1,3	Expensive
8	Toyota Agya	2016	Rp200.000.000,00	1,3	Expensive
9	Toyota Agya	2020	Rp136.000.000,00	1,3	Expensive
10	Toyota Avanza	2020	Rp152.000.000,00	1,3	Expensive

From these calculations, 4 data are included in the class category labeled Cheap and the other 9 are included in the Expensive category.

3. 3 Accuracy Calculation

From these calculations, 4 data are included in the class category labeled Cheap and the other 9 are included in the Expensive category. The following is an accuracy calculation based on the results of the previous classification process.

Table 9. Accuracy Result

Car	Accuracy Data			
	Training Data Class Value	Test Data Class Value	Correct Prediction	Wrong Prediction
1	Cheap	Cheap	1	0
2	Cheap	Cheap	1	0
3	Cheap	Cheap	1	0
4	Cheap	Expensive	0	1
5	Cheap	Cheap	1	0
6	Cheap	Expensive	0	1
7	Expensive	Expensive	1	0
8	Expensive	Expensive	1	0
9	Expensive	Expensive	1	0
10	Expensive	Expensive	1	0
		Total	8	2
		***Accuracy or Inaccuracy* (%)**	80	20

3. CONCLUSION

This research successfully uses the Naive Bayes algorithm to predict used car purchasing decisions based on a dataset that includes attributes such as brand, year of production, mileage, engine condition, completeness of features, and maintenance history. This algorithm utilizes a probabilistic approach to determine the probability of a transaction falling into the "Expensive" or "Cheap" category. By applying the Gaussian Naive Bayes method to continuous attributes such as price, this research succeeded in increasing accuracy in analyzing complex data. Test results show that the prediction model achieves an accuracy of 80%, with the majority of class predictions carried out correctly. This level of accuracy shows the ability of the Naive Bayes algorithm to solve data classification problems. Additionally, further analysis revealed that attributes such as price, model, and engine capacity have a significant influence on purchasing decisions.

ACKNOWLEGDEMENTS

During the planning process, conducting research, and completing this journal, our team faced various challenges. With great gratitude, we acknowledge that this success cannot be separated from the grace and help of God Almighty. We would like to express our sincere respect and gratitude to Mrs. Fitri Marisa, S.Kom., M.Pd., Ph.D., for the guidance, support and encouragement provided throughout the journey of this research. His dedication and thinking have been a big boost to the progress of our research. We also express our appreciation to all fellow researchers and other parties who played a role in the process of data collection, analysis and preparation of this journal. The support and cooperation that exists really helps us in achieving optimal results.

REFERENCES

[1] L. Raheja, “Statistical Analysis of Used Cars,” vol. 4, no. 2, pp. 1115–1122, 2024.
[2] Z. Sun, “Research on factors affecting second-hand car market prices,” *Theor. Nat. Sci.*, vol. 36, pp. 128–135, 2024, doi: 10.54254/2753-8818/36/20240532.
[3] C. Li, “Machine Learning-Based Models for Accurate Car Prices Prediction,” *Highlights Business, Econ. Manag.*, vol. 40, pp. 416–421, 2024, doi: 10.54097/9zcpv779.
[4] A. S. Shaik, N. Shaik, and C. K. Priya, “International Journal of Current Science Research and Review Predictive Modeling in Remote Sensing Using Machine Learning Algorithms,” vol. 07, no. 06, pp. 4116–4123, 2024, doi: 10.47191/ijcsrr/V7-i9-39.
[5] A. Ravi, M. Surabhi, and C. Shah, “Machine Learning Applications in Predictive Maintenance for Vehicles: Case Studies,” *Int. J. Eng. Comput. Sci.*, vol. 11, pp. 25628–25640, 2022, doi: 10.18535/ijecs/v11i08.4707.

[6] H. Mustapha, M. Birjali, and A. beni hssane, "Used Car Price Prediction using Machine Learning: A Case Study," 2022, pp. 1–4. doi: 10.1109/ISIVC54825.2022.9800719.

[7] B. E. Putro and D. Indrawati, "Data Mining Analytics Application for Estimating Used Car Price during the Covid-19 Pandemic in Indonesia," *J. Ilm. Tek. Ind.*, vol. 21, no. 2, pp. 149–161, 2022, doi: 10.23917/jiti.v21i2.18975.

[8] L. Bukvić, J. Škrinjar, T. Fratrović, and B. Abramović, "Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning," *Sustainability*, vol. 14, p. 17034, 2022, doi: 10.3390/su142417034.

[9] F. Ibna Rahman, A. Hasnat, and A. Lisa, "Traffic flow prediction by incorporating weather information in Naïve Bayes Classifier," *J. Adv. Civ. Eng. Pract. Res.*, vol. 8, pp. 10–16, 2019.

[10] R. Ndung'u, R. Ndung'u, and G. Wambugu, "Developing Hybrid-Based Recommender System with Naïve Bayes Optimization to Increase Prediction Efficiency," *Int. J. Comput. Inf. Technol.*, vol. 10, 2021, doi: 10.24203/ijcit.v10i2.75.

[11] Y. Zhu, "Prediction of the price of used cars based on machine learning algorithms," *Appl. Comput. Eng.*, vol. 6, pp. 785–791, 2023, doi: 10.54254/2755-2721/6/20230917.

[12] J. Paul, "Predictive Analytics in the Automobile Industry: Driving Efficiency and Customer Satisfaction," 2022.