The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4147

Analysis of Consumer Preferences for Used Cars Using K-Means Clustering as a Basis for Building Business Actions

Feri, Endang Amalia[2], Ucu Nugraha[3], Ahmad Arif[4]
feri.sulianta@widyatama.ac.id[1], endang. amalia@widyatama.ac.id[2],
ucu.nugraha@widyatama.ac.id[3], ahmad8117@widyatama.ac.id[4]
[1,2,3,4]Faculty of Engineering Widyatama University

Article Information

Received : 17 Jun 2024
Revised : 26 Jul 2024
Accepted : 8 Aug 2024

Keywords

Data Mining, Used Cars, Consumer Preferences, K-Means Clustering,

Abstract

Transportation is a crucial aspect of modern mobility, enabling the movement of goods and individuals. Private vehicles, such as motorcycles and cars, are the dominant choice for many due to their convenience and practicality. However, with numerous car options, consumers often struggle to select used cars that match their needs and preferences. In this research, data mining is employed to cluster used cars based on consumer preferences for price, rating, and performance using the K-Means Clustering algorithm. The data consists of used car information from various sources. The K-Means algorithm groups the cars into clusters based on price, rating, and performance criteria. The results reveal three consumer groups with different preferences in these areas. This information provides valuable insights for consumers to make informed decisions and offers guidance to the automotive industry for designing more effective business actions.

A. Introduction

Transportation is a useful tool for moving goods or people in specific quantities, from one place to another, and within a certain time frame[9]. The need for transportation has become a primary need, replacing public transportation for some people [2][3]. The main reason people prefer private vehicles, especially motorcycles or cars, is because they are considered more comfortable, luxurious, and practical for both city and intercity travel [8]. With the advancement of time, more car options are offered by manufacturers. This is accompanied by intense advertising of new cars, which makes some consumers interested and driven to trade (sell) their old cars and replace them with the latest models. This phenomenon creates a market for used cars that are still feasible and can be sold to other consumers [7].

Choosing a used car is an important decision for consumers who want to buy a vehicle [14]. Used cars are often a more affordable option than new ones, but consumers often face difficulties in selecting a used car that suits their needs and preferences [17]. Various factors such as price, rating, and performance of the car are the main considerations in the selection process of a used car [6].

The K-Means algorithm is one of the methods in data analysis used to cluster data into several groups based on their characteristics. In the context of analyzing consumer preferences for used cars based on price, rating, and performance, the use of K-Means Clustering allows the grouping of consumers into clusters based on their preferences for these factors [1][11][12][13].

By applying K-Means Clustering to consumer preference data for used cars, the study can identify patterns or clusters of consumers with similar preferences regarding price, rating, and performance of used cars. This allows for a deeper understanding of consumer preferences and provides valuable insights for the automotive industry in designing more effective marketing strategies targeting consumer preferences more specifically [19][20].

B. Research Method

Data Mining

The CRISP-DM (Cross-Industry Standard Process for Data Mining) method is a standard process used for data mining projects. Here are the steps that can be followed for analyzing consumer preferences for used cars based on price, rating, and performance using K-Means Clustering [5]:

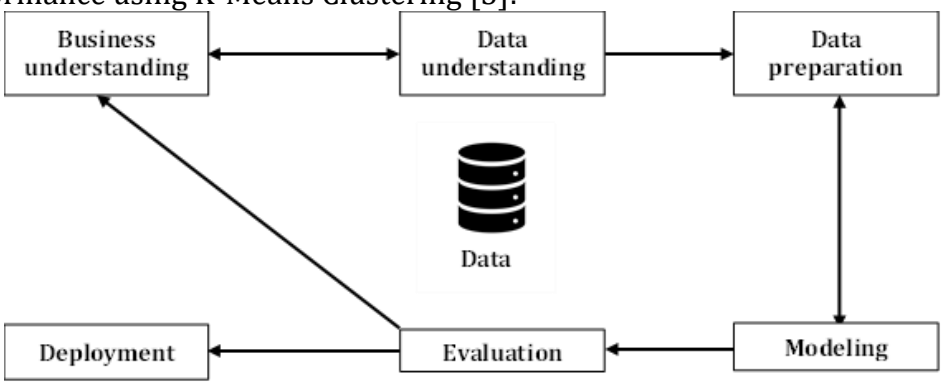

Figure 1. The CRISP-DM (Cross-Industry Standard Process for Data Mining) process

In the CRISP-DM process for analyzing consumer preferences for used cars based on price, rating, and performance using K-Means Clustering, each phase plays a crucial role as below:

- Business Understanding: This phase involves understanding the business goals and needs, focusing on consumer preferences for used cars. With K-Means Clustering, the data mining objective is to cluster consumers based on their preferences regarding price, rating, and performance to enhance targeted marketing strategies.
- Data Understanding: In this phase, data is collected from various sources such as used car sales websites and consumer reviews. Data quality verification is performed to identify errors, missing values, inconsistencies, and uncover potential hidden information for further hypothesis development.
- Data Preparation: Data preparation involves cleaning the data by handling missing values, correcting errors, and normalizing or standardizing features to prepare raw data for clustering modeling.
- Modeling: The modeling phase employs the K-Means algorithm to cluster consumers based on their preferences for price, rating, and performance. Techniques like the Elbow Method or Silhouette Score are used to determine the optimal number of clusters, and model parameters are adjusted for optimal results.
- Evaluation: Model evaluation ensures its quality from both a data analysis perspective and achievement of initial objectives. Internal evaluation metrics such as inertia and Silhouette Score assess clustering quality, while external evaluation examines the usefulness of clustering results in a business context.
- Deployment: The deployment phase involves converting acquired knowledge into a usable format for stakeholders. The model is integrated into production systems to develop marketing strategies and tailor product offerings based on consumer preferences. Performance monitoring and model updates are conducted as needed to ensure ongoing relevance.

This structured approach ensures systematic exploration and application of data mining techniques to enhance decision-making and strategy formulation in the used car market.

K-Means Clustering

K-Means Clustering divides data based on similarity in features such as price, rating, and performance of used cars. The process involves determining the number of clusters kk, selecting random centroids, assigning data points to the nearest centroid, and updating centroids until stability is achieved. Clustering results help understand consumer preferences and develop effective marketing strategies [5][15][16].

Clustering is a formal study for partitioning or grouping data without using predefined category labels. It is an unsupervised learning method without data training stages, unlike classification. Clustering is used to naturally group data based on similarities among data objects while minimizing similarity to other clusters [5].

The research design used in this study, as depicted in Figure 1 below [14], includes the implementation details and visual representation of the research methodology.

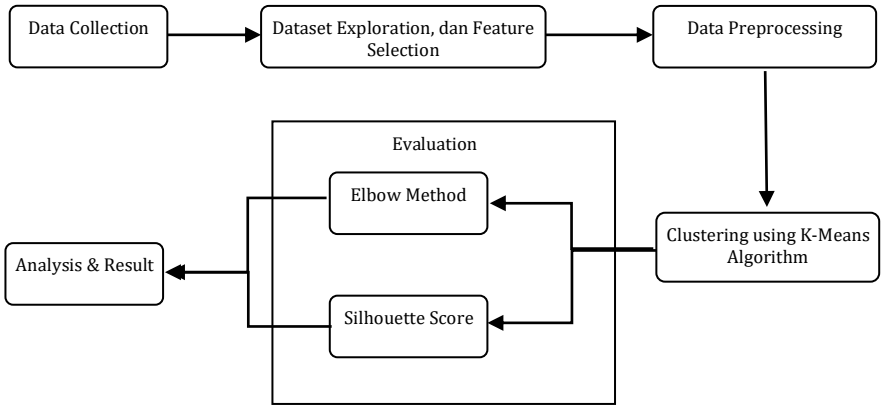

Figure 2. Diagram of Research Design

Data Mining Process

The car data used in this study was obtained through a collection process involving scraping postings from popular online car marketplaces, under the dataset name "CarsForSale". This dataset contains comprehensive information about various used cars being sold online, including attributes such as brand, model, year of manufacture, exterior color, and more. The data collection process utilized scraping techniques, which involve automatically extracting information from web pages of popular online car marketplace websites. The data used consists of vehicle data from various car brands produced from 2001 to 2022, with 9380 records.

The data exploration stage of "CarsForSale" involved examining attributes to understand data characteristics and distributions. Two tables were used: one for initial column descriptions and another for key columns in the K-Means Clustering analysis, including price, consumer rating, performance, comfort, and other factors influencing consumer preferences.

Table 1. Samples of data that contain missing values

Year	Make	Model	Used/New	Price	Consumer Rating
2019	Honda	Civic Sport	Used	Not Priced	4.9
2019	BMW	540 i xDrive	Used	Not Priced	4.9
2019	Jeep	Cherokee Limited	Jeep Certified	Not Priced	4.6
2019	Honda	Civic Sport	Used	Not Priced	4.9
2020	Honda	CR-V LX	Used	Not Priced	4.8

Data preprocessing for the analysis of consumer preferences for used cars using K-Means Clustering aims to ensure consistent and structured data. Steps

include handling outliers, missing values, and data duplicates, as well as transforming categorical data into numeric and scaling if necessary. In the "CarsForSale" dataset, there are no data duplicates, but there are 5 entries in the "Price" column with the value "Not Price" which is non-numeric, indicating missing values that need to be removed before analysis can proceed. The selected dataset for preference analysis using K-Means Clustering needs to be normalized by converting the data in the "Price" column into numeric form.

In conducting this test, the K-Means method is employed to determine clusters in the processed data. Cluster formation using the K-Means method involves determining the value of kk as the number of clusters to be formed, attempting to initialize with 2 clusters in the K-Means algorithm with output in Jupyter. Next, identifying the k-point or centroids randomly, as well as the centroids obtained from the "CarsForSale" dataset, and finally calculating the distance of each input data from each centroid using the Euclidean Distance formula until the closest distance from each data to the centroid is found. The results of the proximity of centroid distances with data from the dataset can be visualized.

Table 2. Centroid of each cluster

Centroid	Ordinate	X Axis
C0	0.9032279	0.93693083
C1	0.62411844	0.77449529

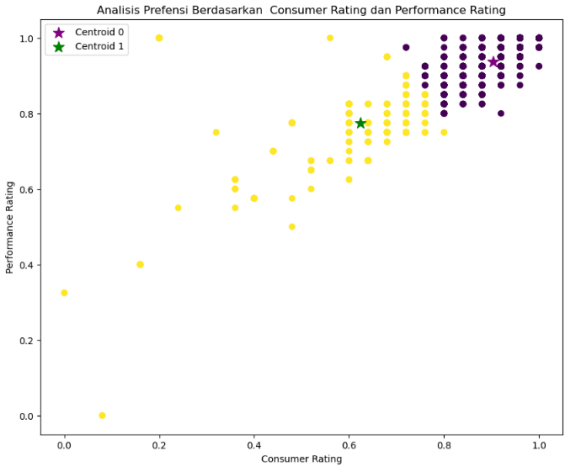

Figure 3. Visualization of the K-Means Clustering results with 2 clusters

Price categorization for cars is created in the "Price" column by adding a new column "Price Category" to group prices based on global market ranges, price aggregators, industry reports, as well as automotive articles and publications, namely: Low Price for prices below $20,000, Middle Price for prices ranging between $20,000 - $50,000, and High Price for prices exceeding $50,000.

Before evaluating the clustering results on the "CarsForSale" dataset using the K-Means algorithm, there are two important evaluation methods to consider. The Elbow Method is used to find the optimal number of clusters by plotting SSE against the number of clusters. In the context of evaluating the "CarsForSale" dataset using the K-Means algorithm with the Elbow Method, it was found that the

optimal number of clusters is 2, as indicated by the "elbow point" in the SSE graph below:

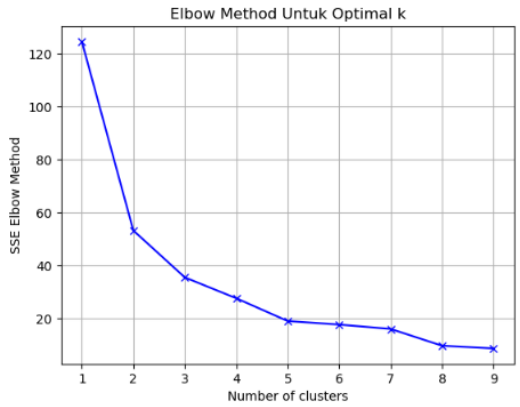

Figure 4. Visualization of the SSE Elbow Method for Optimal K

The Silhouette Score is used to evaluate the quality of cluster separation. In evaluating using the Silhouette Score on the "CarsForSale" dataset with the K-Means algorithm, it was found that the highest Silhouette Score occurs when the number of clusters is 2, indicating that these clusters exhibit good separation between clusters and high cohesion within clusters in the figure below:

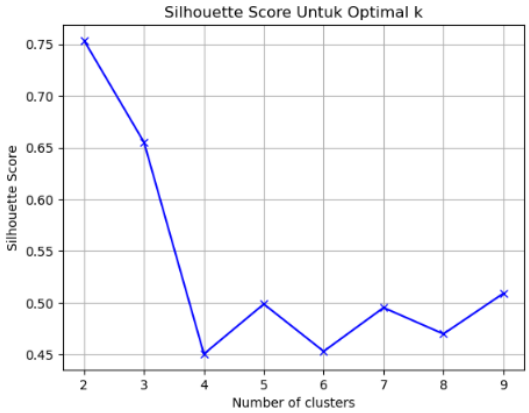

Figure 5. Visualization of the Silhouette Score for Optimal K

Based on the evaluation results using the Elbow Method and Silhouette Score on the "CarsForSale" dataset with the K-Means algorithm, it was found that the optimal number of clusters is 2. These findings indicate that the used cars data can be effectively divided into two main groups with good separation between clusters and high cohesion within clusters. These clusters likely reflect different price characteristics or varying consumer preferences for used cars. Therefore, this clustering analysis can contribute to a better understanding of consumer preferences and dynamics in the used car market, enabling the development of more targeted marketing strategies and better product adjustments according to market needs.

C. Result and Discussion

The results of this study categorize used cars based on price and performance. Cars are divided into three price categories, suach as High Price with Cars priced at $50,000 or higher, Middle Price with Cars priced between $20,000 and $50,000, and Low Price with Cars priced below $20,000. Performance assessment of cars is based on ConsumerRating and PerformanceRating (ranging from 1.0 to 5.0). The total rating score is calculated as the sum of both rating columns, such as Good, which is Cars with a total rating score of 8.9 or higher and Less Good, which is Cars with a total rating score less than 8.9.

Below are the results of the analysis showing the number of cars in each price category and performance cluster:

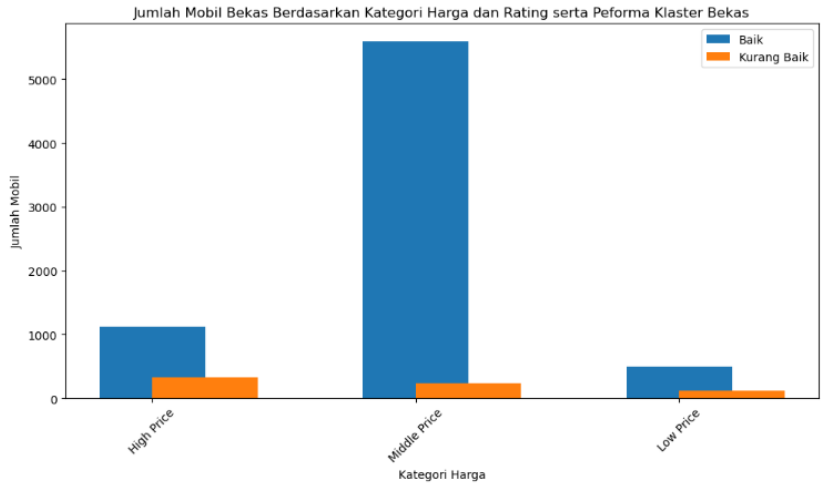

Figure6. Visualizing Bar Chart of Number of Used Cars Based on Price Categories and Ratings as well as Performance Cluster

And to obtain more detailed data from the bar chart visualization of the Number of Used Cars Based on Price Categories and Ratings as well as Performance Cluster in the table below:

Table 3. Details of the Number of Used Cars Based on Price Categories and Ratings as well as Performance Cluster

Condition	Total Cars	High Price Cars	Middle Price Cars	Low Price Cars
Good	7225	1127	5595	503
Less Good	671	326	231	114

D. Conclusion

This study provides insights for consumers of used cars through data analysis and visualization that categorizes cars based on price and performance. It helps consumers manage expectations, select cars with the best value, and optimize their budget. This analysis also reduces purchasing risks by providing clear comparisons between price categories and car performance. The results indicate that many cars in the 'Middle Price' category offer good performance, giving consumers confidence to make wiser and informed decisions.

Based on the findings and conclusions of this research, several recommendations can be made for future studies. Developing more sophisticated

tools for analysis and visualization could assist consumers more efficiently in selecting used cars. Future research could focus on developing more accurate predictive models to forecast the performance of used cars based on historical data, and explore other factors such as age, mileage, and physical condition to provide more comprehensive guidance. Additionally, further studies could explore the use of advanced data mining techniques, such as other clustering algorithms or machine learning methods, to improve the accuracy and relevance of analysis results. Researchers could also evaluate the impact of data visualization on consumer decision-making processes, and explore the potential integration with e-commerce platforms for used cars to provide more personalized recommendations.

E. Acknowledgment

We are immensely grateful to LP2M Widyatama University and Widyatama University for their generosity in providing us with a conducive working environment that supported the progress and success of our research. Their support and resources were crucial to our endeavors, enabling us to delve into this study and make significant contributions to the field.

F. References

[1] A.K. Jain, M.N. Murty, P.J. Flynn, Data Clustering: A Review. ACM Computing Surveys, 32(3), pp. 264-323. 1999.

[2] Anshor, A. H., & Rahardjo, S. B, Analisis Resiko Transformasi Mobil Berbahan Minyak ke Mobil Listrik Dengan Decision Tree. Bulletin of Information Technology (BIT), 4(4), pp. 414-418. 2023.

[3] Anshor, A. H., & W. Wiyanto, Analisis Pembelian Mobil Listrik Menggunakan Metode Analytical Hierarchy Process (AHP) dan Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). KLIK: Kajian Ilmiah Informatika dan Komputer, 4(1), pp. 476-485. 2023.

[4] Atim, S. B, Penerapan Metode Multi-Attributive Border Approximation Area Comparison Dalam Rekomendasi Pemilihan Mobil Second. Journal of Information Technology, Software Engineering and Computer Science, 2(2), pp. 99-110. 2024.

[5] Cheng, C. H., & Chen, Y. S, Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 36(3), pp. 4176-4184. 2009.

[6] Diana Yusuf, Ellya Sestri, Fahrul Razi, Implementasi Teknik Clustering Untuk Pengelompokan Mobil Bekas Berdasarkan Grade Pada Mobi Auto. Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD, 6(2), pp. 307-313. 2023.

[7] Gusrianty, Dwi Oktarina, Wahyu Joni Kurniawan, Sistem Pendukung Keputusan Dengan Metode Promethee Untuk Menentukan Kepuasan Pelanggan Penjualan Sepeda Motor Bekas. Jurnal SISTEMASI, 8(1), pp. 62-69. 2019.

[8] Irawan Setiadi, Sistem Pendukung Keputusan Pemilihan Mobil Bekas Dengan Metode AHP dan SAW Pada Nava Sukses Motor. Jurnal STRING (Satuan Tulisan Riset dan Inovasi Teknologi), 3(3), pp. 247-257. 2019.

[9] Khairul Winata, Sistem Pendukung Keputusan Seleksi Motor Baru Berbasis Android Dengan Metode Simple Additive Weighting. JATI (Jurnal Mahasiswa Teknik Informatika), 1(1), pp. 739-746. 2017.
[10] Lusinia, S. A., Yanto, H., & M. Saputra, Penerapan Metode Weighted Product dalam Memperdiksi Permintaan Konsumen di Salon Mobil Singgalang Motor. Jurnal Teknologi, 10(1), pp. 46-51. 2020.
[11] Mayra Z. Rodriguez, Cesar H. CominID, Dalcimar Casanova, Odemir M. Bruno, Diego R. Amancio, Luciano da F. Costa, Francisco A. Rodrigues, Clustering algorithms: A comparative approach. PLOS ONE, 14(1), pp. 1-34. 2019.
[12] Moufok, S., Mouattah,S., H,Khalid.,K-means and DBSCAN for look-alike sound-alike medicines issue.International Journal of Data Mining, Modelling and Management, 2024 Vol.16 No.1, pp.49 - 65. doi: 10.1504/IJDMMM.2024.136215.
[13] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, pp. 63-67, 2010, doi: 10.1109/IITSI.2010.74.
[14] Simanjuntak, D. P., Priyanto, S., & Irawan, M. Z., Analisis Sensitivitas Pemilihan Moda Bus Damri, Taksi Online Dan Mobil Pribadi Ke Kereta Api Bandara. Journal of Comprehensive Science (JCS), 2(7), pp. 2199-2212. 2023.
[15] Sisilia Fhelly Djun, I Gede Aris Gunadi, Sariyasa Sariyasa. Analisis Segmentasi Pelanggan pada Bisnis dengan Menggunakan Metode KMeans Clustering pada Model Data RFM. JTIM: Jurnal Teknologi Informasi Dan Multimedia, 5(4), pp. 354-364. 2024.
[16] Sulianta, F. Customer Profiling pada Supermarket Menggunakan Algoritma K-Means dalam Memilih Produk Berdasarkan Selera Konsumen dengan Daya Beli Maksimum. Jurnal Ilmiah Teknologi Informasi Terapan. ISSN.2407-3911 Volume 1 No.1, December 2014.
[17] T. Soni Madhulatha. An Overview On Clustering Methods. IOSR Journal of Engineering. Apr. 2012, Vol. 2(4) pp: 719-725.
[18] Untung Surapati, Magfiratul Jannah. Penerapan Data Mining Menggunakan Metode K-Means Untuk Mengetahui Minat Customer Dalam Pembelian Merchandise Kpop. Jurnal Sains dan Teknologi, 5(3), pp. 875-884. 2024.
[19] Yang, J.; Wang, J. Tag clustering algorithm lmmsk: Improved k-means algorithm based on latent semantic analysis. J. Syst. Electron 28, 374–384. 2017.
[20] Yuli Murdianingsih, Ulfah Sitiumayah. Sistem Rekomendasi Pembelian Sepeda Motor Bekas Menggunakan Metode Fuzzy Tahani. Jurnal ICT : Information Communication & Technology, 19(1), pp. 17-26. 2020.