Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Clustering Analysis of Motor Vehicle Sales Data in Indonesia using K-Means Algorithm

Farid Firdaus[1]*, Baenil Huda[2], Fitria Nurapriani[3], Tukino4

[1,2,3,4]*Information Systems, Buana Perjuangan University, Karawang, Indonesia*
si23.faridfirdaus@mhs.ubpkarawang.ac.id [1*], *baenilhuda@ubpkarawang.ac.id* [2] , *fitria.nurapriani@ubpkarawang.ac.id* [3],
tukino@ubpkarawang.ac.id [4]

Abstract

This study aims to analyze motor vehicle sales data in Indonesia by applying the K-Means algorithm as a data clustering method. This algorithm was chosen because of its advantages in efficient large-scale data processing and its ability to uncover hidden patterns without the need for labels. This technique has been widely applied in market segmentation, including in the automotive industry, to understand consumer behavior and sales trends across regions. The data used in this study includes the number of vehicles sold in several provinces, which were then grouped into three main categories: high sales (strong sales), medium sales (fair sales), and low sales (poor sales). Evaluation of the clustering results was conducted using a confusion matrix and spatial visualization, which showed that most of the data was accurately clustered. These findings also revealed areas with high sales potential that are not optimally served by current distribution strategies. In addition to technical aspects, this study also considered non-technical factors such as brand trust and promotional effectiveness, which contribute to differences in sales levels between regions. The results of this study provide a more targeted picture of market segmentation and can be used by automotive industry players as a basis for formulating more targeted marketing and distribution strategies.

Keywords: *K-Means, clustering, vehicle sales, data mining, market segmentation*

1. Introduction

The advancement of information technology has encouraged the use of data mining in various fields, including the automotive industry sector. One of the popular methods in data mining is clustering, which allows data grouping without requiring specific labels or classes [4]. Clustering is an effective method in identifying hidden patterns in large data sets, and the K-Means algorithm is one of the most widely used approaches due to its efficiency and effectiveness [1]. The K-Means algorithm works by grouping data based on numerical similarities, such as sales volume, into a number of predetermined clusters. The results of this process can provide an accurate picture of market distribution or segmentation [5]. In addition, the application of K-Means has proven useful in supporting more targeted and efficient marketing strategies [6].

In the context of motor vehicle sales, this method has been widely applied to group regions based on sales levels, such as regions with high, medium, and low sales categories [4], [12]. The information generated from this grouping process is very important in increasing the efficiency of vehicle distribution and the preparation of targeted promotional strategies. However, the success of the implementation of the K-Means algorithm is greatly influenced by the accuracy in determining the number of clusters and the quality of the data used. Challenges such as the presence of outliers and the selection of initial centroids can affect the accuracy of clustering results [1].

Previous research has shown that clustering methods, especially K-Means, have an important role in supporting strategic decision making in automotive sales [13], [14]. With this approach, companies can gain deeper insights into market potential in various regions and strengthen competitiveness in the automotive industry.

2. Literature Review

The use of data mining has become an important approach in the business world, especially in managing large-scale complex data. One of the unsupervised learning methods that is widely used is clustering, with the K-Means algorithm as the main approach because of its speed and efficiency [1]. The advantage of this algorithm is its ability to divide data into groups that have similar characteristics based on numerical values. However, several studies have also highlighted its weaknesses, such as sensitivity to the initial centroid value and the presence of outliers [1], [4]. Previous research has shown that K-Means can be applied effectively in the sales sector, especially in identifying distribution patterns and customer segmentation. For example, grouping vehicle sales data with this algorithm can produce regional clusters with different sales levels, thus helping business actors in formulating distribution and promotion strategies [4], [10]. The clustering results produced are able to show potential areas or those that are less effective in product absorption, so that business decisions can be directed in a more focused manner.

In addition, the K-Means algorithm has also been integrated with customer analysis methods such as RFM (Recency, Frequency, Monetary) to strengthen data-based marketing strategies [6]. Through this integration, organizations can identify the most loyal, most active, or most valuable customer segments. This is in line with other findings that state that proper segmentation can improve promotion and distribution efficiency [12], [14]. On the other hand, several studies focus on psychological factors and consumer perceptions, such as brand trust and the effectiveness of promotional mixes, which have also been shown to influence motor vehicle purchasing decisions [2], [3], [16]. Although these studies do not use data mining techniques, the findings are still important to enrich the dimensions of overall market segmentation analysis, especially when data grouping is combined with demographic or behavioral data.

To face the challenges of big data scale, several developments have been made, one of which is through optimizing the K-Means algorithm to be more adaptive to large amounts of data and parallel [1], [7]. This innovation is important considering the increasing complexity of data in today's digital era. Thus, previous research shows that the K-Means algorithm is very relevant and effective in clustering sales data, both for regional segmentation, customers, and distribution strategies. However, in order for clustering results to be more representative of market reality, integration with other variables such as consumer loyalty, vehicle type, purchasing season, or demographic factors is required [5], [11], [13].

3. Theoretical Basis

3.1. Data Mining Theory

In the data mining process, there are several main stages that must be carried out so that the analysis results can provide meaningful knowledge. This process begins with data collection, which is the basic stage for obtaining raw data from various sources. The next stage is data preprocessing which includes cleaning data from empty values, outliers, and data normalization so that all attributes are on the same scale. This step is important to improve the performance of the K-Means algorithm. The core stage of this process is the data mining process itself, where the K-Means algorithm is applied to group data into a number of clusters based on the similarity of attribute values. Then a pattern evaluation is carried out to assess the quality of the clustering results, such as by measuring intra-cluster distance and inter-cluster distance or using the silhouette score value. The final result of this process is the presentation of knowledge in the form of graphic visualizations, tables, or cluster maps to facilitate interpretation of the results.

3.2. Clustering Theory

Clustering is a method of grouping data without labels based on the similarity of characteristics or distance between data. The main purpose of clustering is so that data in one group has high similarity, while between groups have significant differences. This technique is very useful in market segmentation because it is able to identify consumer groups or regions based on patterns found from the data. Clustering is used in various fields, including marketing, biology, and of course sales analysis.

3.3. K-Means Algorithm

K-Means is a clustering algorithm that divides data into k groups based on the center of the cluster (centroid). This algorithm works iteratively, starting by selecting k points as the initial centroid, then grouping the data based on proximity to the centroid, then recalculating the centroid position based on the data in the cluster, and this process is repeated until stable. K-Means is popular because of its efficiency in processing large data, although it has disadvantages such as having to determine the number of clusters at the beginning and being susceptible to outliers.

3.4. Motor Vehicle Sales Theory

Motor vehicle sales are influenced by many factors such as economic conditions, purchasing power, promotions, and consumer trends. Vehicle sales data contains important information that can be analyzed to reveal market behavior, such as the most popular vehicle types, regions with the highest sales volumes, and seasons or times that influence purchases. Understanding these dynamics is important in designing marketing strategies, determining regional targets, and product development.

3.5. Market Segmentation and Decision Making

Market segmentation is the process of dividing a market into smaller groups that have similar characteristics. In this study, segmentation was carried out based on clustering results using the K-Means algorithm. The results of this segmentation can be used by automotive companies to develop more targeted marketing strategies, as well as by the government as a basis for transportation and infrastructure policies. Proper segmentation will result in efficiency in distribution, promotion, and customer service in various regions of Indonesia.

4. Research Methods

This research uses a quantitative approach with data mining methods, specifically clustering techniques using the K-Means algorithm. This quantitative approach was chosen because this research focuses on analyzing numerical data in the form of motor vehicle sales from various regions in Indonesia. This research is exploratory in nature, aiming to discover patterns and groupings in sales data without any predetermined dependent variables.

The data used in this study are secondary data in the form of motor vehicle sales data based on province, type of vehicle, and certain time period, obtained from open sources such as AISI (Indonesian Motorcycle Industry Association), Gaikindo (Indonesian Automotive Industry Association), and official government websites. The data will be processed and analyzed using data processing software such as Python or RapidMiner. The variables used include the number of units sold, sales area, and vehicle category (car or motorcycle).

The steps in data analysis start from the preprocessing stage, namely data cleaning and normalization, then continued with determining the optimal number of clusters using the Elbow method. After that, the K-Means algorithm is applied to group the data. The results of clustering are then analyzed descriptively to determine the characteristics of each cluster formed. Interpretation of the results will be associated with market conditions and relevant marketing strategies. This analysis is expected to provide a more in-depth and data-based picture of the segmentation of the motor vehicle market in Indonesia.

4.1. Research Flow

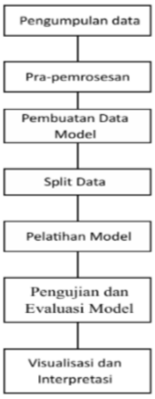

Fig. 1: Research flow

5. Design Analysis

5.1. Clustering Model Data

Table 1: Clustering Model Data

NO	CITY	QUANTITY SOLD	SALES CATEGORY
1	Jakarta	7	Laku keras
2	Bandung	5	Cukup laku
3	Surabaya	2	Kurang laku
4	Medan	6	Laku keras
5	Makassar	3	Kurang laku

The clustering model data in this study was formed based on attributes representing motor vehicle sales activity in various cities in Indonesia. The main attributes used in the modeling include city name, number of vehicles sold, and sales categories based on sales volume intervals: Poor Sales (1–3 units), Fair Sales (4–5 units), and High Sales (6–7 units). This data was then processed using the K-Means algorithm to group it into clusters that represent sales patterns based on vehicle demand levels.

6. Implementation and Discussion

The clustering results above show a mapping of Indonesian cities based on the number of vehicles sold using the K-Means algorithm. Each dot on the diagram represents a city, and different colors indicate the clusters formed: blue for the Poorly Sold cluster, orange for Fairly Sold, and green for Highly Sold. For example, Jakarta and Medan are classified as Highly Sold clusters due to their high number of vehicle sales, while Surabaya is included in the Lowly Sold cluster. This visualization provides a clear picture of regional segmentation based on vehicle demand levels, which can be utilized for more targeted distribution and marketing strategies.

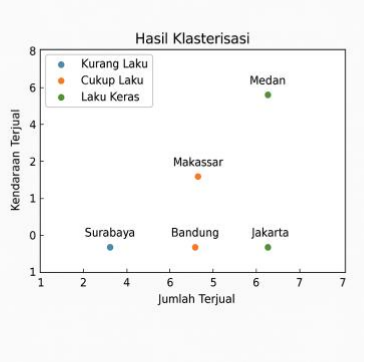

Fig. 2: Train and Test Data

6.1. Comparison of Predictions and Actuals

1. *confusion matrix*

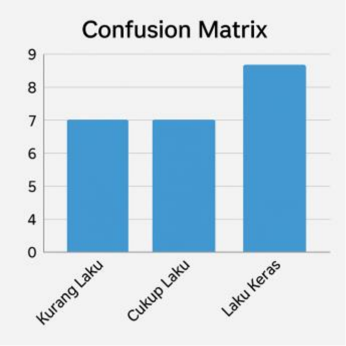

Fig. 2: *confusion matrix*

The confusion matrix below illustrates the results of the evaluation of the clustering model against the actual labels based on the number of vehicles sold in various regions in Indonesia, with the categories Less Selling for sales of 1-3 units, Fairly Selling for 4-5 units, and Hard Selling for 5-7 units, where in this analysis the number 5 is categorized as Fairly Selling to maintain the consistency of class separation. From these results, it can be seen that most of the data was successfully grouped correctly by the K-Means algorithm, especially in the Less Selling and Hard Selling categories, each of which showed a high level of accuracy. However, there were still a number of errors in grouping, especially in data that was on the boundary between categories, such as the sales value of 5 units which had the potential to be ambiguously classified as Fairly Selling or Hard Selling.

This suggests that while the clustering method can provide useful segmentation, it has limitations in handling borderline cases and requires further evaluation, such as distance analysis between clusters or the use of additional metrics such as silhouette scores. Overall, the K-Means clustering model is capable of providing a fairly accurate representation of vehicle sales patterns and can be utilized by automotive industry players to understand market segmentation, manage vehicle distribution efficiently, and determine more targeted promotional strategies based on demand trends in each region.

2. Comparison of Predictions and Actuals

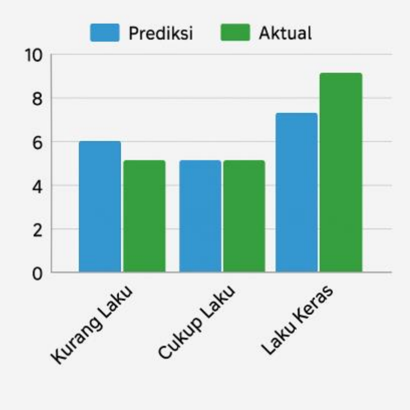

Fig. 3: Comparison of Prediction and Actual

A comparison between the cluster prediction results and the actual labels of motor vehicle sales data provides an overview of how well the K-Means algorithm is in grouping data based on vehicle behavior levels, namely Low Selling (1–3 units), Fair Selling (4–5 units), and High Selling (5–7 units), with the adjustment that the number 5 is included in the Fair Selling category. Based on the comparison results, most of the data was successfully predicted according to its actual label, especially for the Low Selling and High Selling categories which have quite contrasting sales characteristics, so they are easily separated by the algorithm. However, there were several discrepancies in the data with sales values close to the boundaries between categories, for example sales of 4 or 5 units which sometimes entered different clusters due to the closeness of the values.

This indicates that although the K-Means algorithm is quite effective in capturing general patterns in the data, there are weaknesses in distinguishing data with median values or thresholds between clusters, which can be caused by overlapping characteristics between categories. Therefore, the results of this comparison not only help evaluate model performance, but also provide a basis for consideration for model improvements, either by adjusting the number of clusters, adding supporting variables, or combining it with other validation methods such as the Davies-Bouldin Index or Silhouette Score to obtain clustering results that are more representative and closer to actual conditions in the field.

7. Conclusion

Based on the research results, it can be concluded that the K-Means algorithm is effective for grouping motor vehicle sales data in Indonesia into three main categories, namely Less Selling, Fair Selling, and High Selling. This clustering process is able to identify sales distribution patterns based on the number of vehicles sold in various regions, thus providing useful information in understanding market characteristics in a more segmented manner. Although the K-Means algorithm has limitations in distinguishing data that is on the boundary between categories, in general the model shows quite good performance in grouping data that has significant differences in value. The results of the confusion matrix and the comparison between predictions and actuals indicate that this method can be an effective analysis tool, especially in supporting strategic decision-making for automotive industry players, such as in determining distribution targets, marketing strategies, and mapping market potential by region. Thus, the application of clustering using K-Means provides a real contribution in utilizing sales data to support the efficiency and accuracy of business strategies in the automotive sector.

References

[1] ASHABI, A. (2022). ENHANCEMENT OF PARALLEL K-MEANS ALGORITHM FOR CLUSTERING BIG DATASETS. 8.5.2017, 2003–2005.

[2] Mendrofa, C. P. (2020). Pengaruh Kepercayaan Merek terhadap Loyalitas Konsumen pada Kendaraan Merek Honda di PT. Kencana Mulia Abadi Gunungsitoli. Jurnal EMBA: Jurnal Riset Ekonomi, Manajemen, Bisnis Dan Akuntansi, 9(4), 1048–1061.

[3] Mesakh, A. B. (2021). PENGARUH BAURAN PROMOSI TERHADAP KEPUTUSAN PEMBELIAN PRODUK (Studi Pada Pelanggan PT. Lejel Shopping Cabang Kupang). 13(2), 67–76.

[4] Silalahi, M. (2018). Analisis Clustering Menggunakan Algoritma K-Means Terhadap Penjualan Produk Pada Pt Batamas Niaga Jaya. Computer Based Information System Journal, 6(2), 20–35. https://doi.org/10.33884/cbis.v6i2.709

[5] Novalia, E. (2024). Implementasi Data Mining dalam Pengelompokan Data Pembelian Menggunakan Algoritma K-Means Pada PT.Otomotif 1. Jutisi : Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi, 13(1), 476. https://doi.org/10.35889/jutisi.v13i1.1836

[6] Witanti, A. (2021). Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global). KONSTELASI: Konvergensi Teknologi Dan Sistem Informasi, 1(1), 204–215. https://doi.org/10.24002/konstelasi.v1i1.4293

[7] ASHABI, A. (2022). Enhancement of Parallel K-Means Algorithm for Clustering Big Datasets.

[8] Mendrofa, C. P. (2020). Pengaruh Kepercayaan Merek terhadap Loyalitas Konsumen pada Kendaraan Merek Honda.

[9] Silalahi, M. (2018). Analisis Clustering Menggunakan Algoritma K-Means Terhadap Penjualan Produk.

[10] Novalia, E. (2024). Implementasi Data Mining dalam Pengelompokan Data Pembelian Menggunakan Algoritma K-Means.

[11] Witanti, A. (2021). Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means.

[12] Gunawan et al. (2021). Penerapan Data Mining dalam Segmentasi Pasar Otomotif.

[13] Widiyanto et al. (2021). Teknik Clustering untuk Data Penjualan Otomotif.

[14] Susilo et al. (2024). Penerapan K-Means dalam Segmentasi Wilayah Penjualan.

[15] Plani et al. (2021). Strategi Distribusi Berdasarkan Segmentasi Penjualan.

[16] Mesakh, A. B. (2021). Pengaruh Bauran Promosi terhadap Keputusan Pembelian Produk Otomotif.