International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 7, No. 2, (2023), pp. 118-124

Application of Data Mining For Clustering Car Sales Using The K-Means Clustering Algorithm

Michael Nico Hutasoit[1], Riska Yanu Fa'rifah[2] and Rachmadita Andreswari[3]
[1,2,3]Universitas Telkom, Bandung, West Java, Indonesia
Email: [1]michaelnicohh@student.telkomuniversity.ac.id,
[2]riskayanu@telkomuniversity.ac.idl, [3]andreswari@telkomuniversity.ac.id

Abstract

In the digital era, data is at the core of business continuity. The need for fast, precise and accurate information is needed. Cars are one of the tertiary needs. This means of transportation is a relatively fast development and innovation business. Car sales in Indonesia recorded a reasonably high number in 2014 - 2018, namely 4.157.580 units sold. The highest sales were MPV car types being the most popular type of car, and there are many types of cars in Indonesia, including Sedans, SUV, 7 Seater SUV, and City Car types, and the enthusiasts need to play more. Hence, it is exciting to study. The variety of car brands with competing prices makes it difficult for consumers to choose the right car to buy according to their needs. This can be solved by applying data mining to cluster car sales using the k-means clustering algorithm. The goal is to know the characteristics of the car from each attribute. The k-Means algorithm is used for cluster formation based on five attributes: CC, Tank Capacity, GVW (Kg), Seater, and Door. The elbow and silhouette score methods determine the optimal number of clusters (k). The result is 4 clusters, cluster 0 (High-Performance Heavy Car), cluster 1 (Small Family Car), cluster 2 (High-Performance Small Car), and cluster 3 (Medium Performance Car). The 4 Cluster results are based on the evaluation/validation of the Elbow Method and Silhouette.

Keywords: *We would like to encourage you to list your keywords in this section*

1. Introduction

In this digital era, data is at the core of business continuity. The need for fast, precise and accurate information is needed. However, the fact is that the high need for information is different from adequate information presentation. Cars are one of the needs that belong to the tertiary class. This means of transportation is a relatively fast development and innovation business [1]. Car sales in Indonesia recorded a relatively high number in the period 2014 - 2018 based on data obtained from GAIKINDO (Gabungan Industri Bermotor Indonesia), namely 4,157,580 units sold, the highest sales were MPV car types being the most popular type of car, there are many types of cars in Indonesia including Sedan, SUV, SUV 7 Seater, and City Car types and the enthusiasts are not playing very much. Hence, it is exciting to study. All GAIKINDO members are brand holder agent (APM) companies consisting of motor vehicle manufacturers, distributors, and principal component makers (manufacturers).

Car sales data in Indonesia shows that the highest sales were in 2014, with 1,208,028 car sales units. Then in the following year, in 2015, sales decreased from the previous year, only 1,013,518 units of cars were sold in that year, then in the following year, 2016, sales were recorded at 1. 062,694 units, experienced a slight increase from the previous year, then two years later although not many sales still increased from the previous year, recorded in 2017 as many as 1,077,365 units were sold and in 2018 as many as 1,151,284 units were sold.

The variety of car brands, such as Honda, Toyota, Suzuki, Daihatsu, BMW, Mercedes Benz, Mitsubishi, etc., with competing prices, makes it difficult for consumers to determine the right car to buy according to their needs. This is where the importance of

ISSN : 2580-7250

applying data mining to cluster car sales using the k-means clustering algorithm. Clustering results can be obtained by processing sales data from the previous month.

Clustering is a method used to divide a series of data into several groups based on predetermined similarities. A cluster is a group or set of data objects that are similar to each other in the same cluster and dissimilar to objects in different clusters. Objects will be grouped into one or more clusters so that objects in one cluster will have high similarity. Researchers use this clustering to discover overall distribution patterns and exciting relationships between data attributes. In data mining, efforts are focused on finding methods to cluster large databases effectively and efficiently. Some of the clustering requirements in data mining include scalability, handling different attribute types, handling high dimensionality, handling data with noise, and being translatable.

Data Mining is the process of extracting information from data sets using algorithms and techniques involving the fields of statistics, machine learning, and database management systems [2]. The extraction process from data sets is a technique used in data mining. Data mining is grouped into several sections based on the tasks performed: description, estimation, prediction, classification, clustering, and association. One of the methods in Data Mining is the K-Means Clustering Algorithm, one of the simple and popular machine learning algorithms used to solve data clustering problems.

The K-Means algorithm is chosen because it has a higher accuracy rate than other clustering algorithms. Previous research shows that the K-Means algorithm produces an accuracy rate of 56%, 25% precision and 60% recall, while the K-Medoids algorithm has an accuracy rate of 14%, 25% precision and 25% recall [3]. In addition, K-Means shows that the Davies Bouldin Index value for K-Means validation is 0.161, and the Davies Bouldin Index value for K-Medoids validation is 0.281. Thus, clustering using the K-Means method results better than the K-Medoids method because it produces a smaller Davies Bouldin Index value of 0.16 [4].

2. Research Methodology

2.1. K-Means Algorithm

The k-means algorithm is an iterative clustering algorithm that partitions a dataset into a predefined number of K clusters. The selection of K data points as initial cluster centers also affect the clustering results.

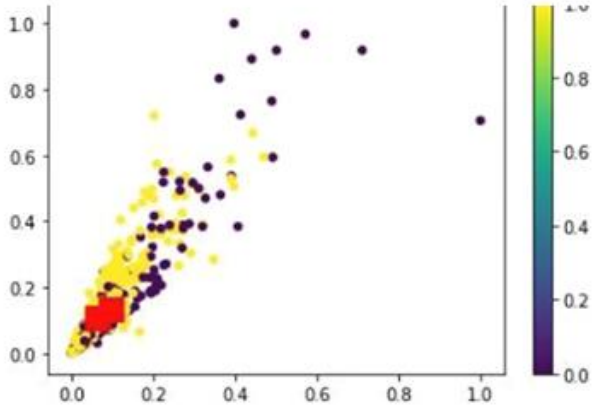

Figure 1. K Means Source: (Purba et al., 2018)

Based on Figure 1, it can be concluded that the natural characteristics of k-means can cause the cluster results obtained in the experiment to differ from those after the clustering process. This condition is known as a local optimum solution, which means that the K-means algorithm is very sensitive to the initial location of the cluster center, where one of the processes of object mining is partitioning an existing object into one or more clusters whose characteristics are similarly grouped in the same cluster.

2.2. Steps of the K-Means Algorithm

The algorithm on K-Means [5-13]:

a. Determine the number of clusters (K), and assign arbitrary cluster centers.
b. Calculate the distance of each data to the cluster center.
c. Group the data into the cluster with the shortest distance.
d. Calculate the cluster center.
e. Repeat steps 2 - 4 until no more data has moved to another cluster.

2.3. Research Dataset

In this study, the data that researchers will study is about car sales data in Indonesia from January to June 2021. The research data source is obtained from Industry Research Data.

Table 1. Research Data

No	1	2	3	4
Sub Brand/Type/Model	Q8 3.0 TFSIA/T	M2 Competition A/T	Grand New Xenia 1.3 XMT	AllNew Sirion 1.3 MT2018
Category	Tipe 4X4	SEDAN	Tipe 4X2	Tipe 4X2
CC	2995	2979	1329	1298
Trans	AT	AT	MT	MT
Fuel	G	G	G	G
Tank Capt	100	0	45	40
GVW (Kg)	2245	2010	1120	930
Gear Ratio	4845	0	5125	4267
Wheel & Tyresize	265/50/R19	245/35 ZR19 93Y; 265/35 ZR19 98Y	185/70R14	175/65R14
PS/HP	272	0	97	90
Wheel Base	3002	2693	2655	2440
Dimension (PxLxT)	5089 x 2002 x 1695	4461 x 1854 x 1410	4190 X 1660 X 1695	3690 X 1665 X 1545
Seater	7	0	7	5
Drive System	4X4	4X2	4X2	4X2
Speed	222	0	0	0
Door	4	0	5	5
Wheels	4	0	4	4
CBU/CKD	CBU	CBU	CKD	CBU
Origin Country	Germany	Germany	INA	Malaysia
Januari	1	3	358	1
Februari	1	3	133	19
Maret	1	1	483	40
April	-	5	353	20
Mei	2	4	235	10
Juni	1	4	708	10
Total	6	20	2270	100

Data Industry Research provides many attributes in the research dataset; here are the attrbutes of the research data.

Table 2. Research Data Attributes

No	Atributte	Information
1	*Sub Brand/Type/Model*	Type/model of vehicle
2	*Category*	Vehicle category
3	CC (*Cubic Centimeter)*	Large capacity of the machine
4	*Trans*	Transmission

5	*Fuel*	Type of vehicle fuel
6	*Tank Capt*	Tank capacity
7	GWV (kg)	Maximum operating weight/mass of the vehicle
8	*Gear Rasio*	The degree of size between the teeth of the transmission
9	*Wheel & Tyresize*	Tire size
10	*Dimension* (PxLxT)	Vehicle dimensions
11	*Seater*	Number of vehicle seats
12	*Drive System*	Car drive system
13	*Speed*	Car speed
14	*Door*	Number of vehicle doors
15	*Wheels*	Number of vehicle tires
16	*CBU/CKD*	Types of cars based on their assembly
17	*Origin Country*	Country of manufacture of the car
18	January	Total sales in January
19	February	Total sales in February
20	March	Total sales in March
21	April	Total sales in April
22	May	Total sales in May
23	June	Total sales in June
24	Totally	Total sales January - June

3. Results and Discussion

3.1. Implementation of the K-Means Algorithm

At this stage, it is done to get the optimal k value of 4. Next, the k-means algorithm will be implemented by clustering 4 clusters. The following code snippet is used for the clustering process using the optimal value (k=4).

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 4, init = 'k-means++', max_iter =
1000, n_init = 100, random_state = 42)
y_kmeans = kmeans.fit_predict(x_scaled)
cluster = pd.DataFrame(y_kmeans)
cluster
```

The following is the result of implementing the k-means algorithm.

Table 3. Cluster Result

CC	Tank Capt	GVW(Kg)	Seater	Door	Cluster
1984	63	1485	5	4	0
1984	54	1490	5	5	2
1984	65	1635	5	4	0
1800	64	1640	5	4	0
…..	…..	…..	…..	…..	…..

```
jumlah_cluster = cluster.value_counts()
jumlah_cluster
```

Table 4. Number of Rows in Each Cluster

Cluster	Jumlah Baris
0	40
1	112
2	55
3	61

Table 4 shows the number of rows representing the data in each cluster that more clusters are formed at cluster index one while the smallest cluster is at cluster index 0.

3.2. Analysis of Cluster Results

Based on the centroid value results, researchers can perform analysis to understand the characteristics of each cluster. The centroid value shows the average attribute value for each cluster after the normalization process. In this context, the normalization process has changed the range of attribute values from 0 to 1.

Table 5. Centroid Points in Each Cluster

Cluster	CC	Tank Capt	GVW(Kg)	Seater	Door	Keterangan
0	0.707465	0.830556	0.663869	0.412500	2.220446	Mobil Berat Kinerja Tinggi
1	0.163938	0.163938	0.198567	0.522321	1.000000	Mobil Kecil Keluarga
2	0.687581	0.512121	0.452554	0.552273	1.000000	Mobil Kecil Kinerja Tinggi
3	0.220397	0.186248	0.383724	0.522541	-2.220446	Mobil Kinerja Menengah

Table 5 shows the centroid value of each cluster.
The following are the characteristics of each cluster based on its centroid value.
1. Cluster 0 (High-Performance Heavy Cars)
This cluster has relatively high centroid values for the CC, Tank Capt, and GVW attributes but has very low centroid values for the Seater and Door attributes. This indicates that this cluster is a group of vehicles with high engine capacity, fuel tank, and weight but very few seats and doors.
2. Cluster 1 (Small Family Car)
This cluster has low centroid values for the CC, Tank Capt, GVW, and Seater attributes but has a maximum centroid value (1) for the Door attribute. This indicates that this cluster is likely a group of vehicles with small size, small fuel tank, light vehicle weight, and five doors. The Door attribute has a value of 1 on the centroid because all vehicles in this cluster have a complete number of doors (5 doors).
3. Cluster 2 (High-Performance Small Car)
This cluster has high centroid values for the CC, Tank Capt, and GVW attributes but has a centroid value of about 0.552273 for the Seater attribute. The Seater value of about 0.552273 indicates that this cluster is likely a group of vehicles with slightly more seats than cluster 1. In addition, this cluster has the maximum centroid value (1) for the Door attribute. This indicates that all vehicles in this cluster have a complete number of doors (5 doors).
4. Cluster 3 (Medium Performance Cars)
This cluster has low centroid values for the CC, Tank Capt, and GVW attributes but has a centroid value of around 6.18 for the Seater attribute. The Seater value of about 6.18 indicates that this cluster is likely a group of vehicles with more seats than Cluster 1 and Cluster 2. However, the value also indicates that this cluster has variation in the number of seats, with some vehicles having seven seats and some having five seats. In addition, this cluster has the maximum centroid value (1) for the Door attribute, but the previous data shows that the Door value in this cluster is 4. This suggests that vehicles in this cluster tend to have four doors.

Based on the results of centroid analysis, it can be concluded that the characteristics of the first cluster include vehicles with large engine capacity, high vehicle weight, and limited number of seats, which are suitable for high performance with a small number of passengers, then for the second cluster includes family vehicles with small size, efficient fuel, and standard passenger capacity, suitable for family daily needs, then for the third cluster includes cars with high performance and more passenger capacity than the first cluster, and finally the fourth cluster includes medium-sized vehicles with a balance between performance and passenger capacity, which are suitable for all-purpose use. The centroid analysis results helped divide the car sales data into four clusters based on their attribute characteristics. Each cluster has unique characteristics and can be identified as a specific vehicle category with different advantages and preferences. The results of this clustering can help consumers make car choices that align with their needs and preferences.

4. Conclusion

From the research that has been done, The results of the implementation of the k-means clustering algorithm on consumer problems in determining the right choice of car to buy according to their needs can be resolved by applying the k-means clustering algorithm. The optimal number of clusters is 4, obtained from testing with the elbow and silhouette score methods. Clustering techniques on consumer problems in determining the right choice of car to buy according to their needs can be resolved by getting the results of four category indices: High-Performance Heavy Cars, High-Performance Small Cars, High-Performance Small Cars, and Medium Performance Cars.

References

[1] Benri, M., Metisen, H., & Latipa, S. (2015). Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila. In *Jurnal Media Infotama* (Vol. 11, Issue 2).

[2] Fitriyadi, A. U. (2021). Algoritma K-Means dan K-Medoids Analisis Algoritma K-Means dan K-Medoids Untuk Clustering Data Kinerja Karyawan Pada Perusahaan Perumahan Nasional. *KILAT*, *10*(1), 157–168. https://doi.org/10.33322/kilat.v10i1.1174

[3] Gunadi, G., & Indra Sensuse, D. (2012). *Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori Dan Frequent Pattern Growth (FP-GROWTH) : Studi Kasus Percetakan Pt. Gramedia* (Vol. 4, Issue 1).

[4] Harahap, F. (2021). *TIN: Terapan Informatika Nusantara Perbandingan Algoritma K Means dan K Medoids Untuk Clustering Kelas Siswa Tunagrahita* (Vol. 2, Issue 4).

[5] Irma T. (2019). Penerapan Data Mining Dalam Dunia Bisnis Menggunakan Metode Clustering. In *Journal Of Institution And Sharia Finance* (Vol. 40).

[6] *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*. (n.d.).

[7] Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI

[8] Nainggolan, R., Perangin-Angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, *1361*(1). https://doi.org/10.1088/1742-6596/1361/1/012015

[9] Nazaruddin, & Sarbaini. (2022). Evaluasi Perubahan Minat Pemilihan Mobil dan Market Share Konsumen di Showroom Pabrikan Honda. *Jurnal Teknologi Dan Manajemen Industri Terapan (JTMIT)*, *1*, 97–103.

[10] Nur Khomarudin, A. (2003). *Teknik Data Mining : Algoritma K-Means Clustering*. https://agusnkhom.wordpress.com
[11] Purba, W., Tamba, S., & Saragih, J. (2018). The effect of mining data k-means clustering toward students profile model drop out potential. *Journal of Physics: Conference Series*, *1007*(1). https://doi.org/10.1088/1742-6596/1007/1/012049
[12] Sutoyo, M. N. (n.d.). *Algoritma K-Means*.
[13] Yanto, R., & Khoiriah, R. (n.d.). *Implementasi Data Mining dengan Metode Algoritma Apriori dalam Menentukan Pola Pembelian Obat*.

Authors

1st Author
Michael Nico Hutasoit
Universitas Telkom, Bandung, West Java

2st Author
Riska Yanu Fa'rifah, S.Si., M.Si
Universitas Telkom, Bandung, West Java

3stAuthor
Rachmadita Andreswari, S.Kom., M.Kom
Universitas Telkom, Bandung, West Java