Journal of Economic Global
Vol. 1 No. 3, May 2024
ISSN: 3026-2941
Journal Homepage: https://penaeducentre.com/index.php/JEcG

Application of Data Mining to Predict Car Sales Using K-Means Clustering Method

Antonia Desya[1)*], Dimas Ilham Pratama[2)], Muchammad Vico Airlangga[3)], Andi Diah Kuswanto[4)]
[1), 2), 3)] Sistem Informasi, Fakultas Teknologi Informasi, Universitas Bina Sarana Informatika
[4)] Sistem Informasi, Fakultas Teknologi Informatika, Universitas Nusa Mandiri

*antoniadesya7@gmail.com

ABSTRACT

The automotive industry is one of the key sectors in the economy. Accurate car sales prediction is crucial for automotive businesses to develop effective marketing and production strategies. Data mining, with its various techniques, offers solutions to assist automotive companies in predicting car sales. One popular data mining technique for predicting car sales is K-Means Clustering. This technique groups car sales data based on characteristics such as car model, price, sales region, and other factors. The clustering results can be used to identify sales patterns and trends, which can then be used to predict future sales. This paper discusses the application of K-Means Clustering for car sales prediction. It explains the steps involved in applying K-Means Clustering, its advantages and disadvantages, and provides an example of its application.

Keywords: Car Sales Prediction, Data Mining, K-Means Clustering Method

INTRODUCTION

The automotive industry is one of the key sectors in the economy (Rezaeinejad, 2021). Accurate car sales prediction is crucial for automotive businesses to develop effective marketing and production strategies. Data mining, with its various techniques, offers solutions to assist automotive companies in predicting car sales. One popular data mining technique for predicting car sales is K-Means Clustering (B. M. Metisen and H. L. Sari, 2015). This technique groups car sales data based on characteristics such as car model, price, sales region, and other factors. The clustering results can be used to identify sales patterns and trends, which can then be used to predict future sales.

Data Mining

Data mining is the process of extracting valuable information from large and complex data sets (H. Annur, 2019). Data mining can be used for various purposes, such as prediction, classification, and clustering (Govindasamy & Velmurugan, 2017; López-Zambrano et al., 2021).

K-Means Clustering

K-Means Clustering is one of the most popular clustering techniques. This technique groups data into k clusters based on the similarity between data points. K-Means Clustering is easy to understand and implement and can handle high-dimensional data.

Car Sales Prediction

Car sales prediction is the process of estimating the number of cars that will be sold in the future. Car sales prediction can be done using various methods, such as regression analysis, machine learning, and data mining (Hutasoit et al., 2023).

The K-Means System

K-Means has the ability to cluster large amounts of data with relatively fast and efficient computation time (K. Fatmawati and A. P. Windarto, 2018). However, K-Means has a drawback caused by the determination of the initial cluster centers (K. Handoko, 2016). The cluster results formed by the K-Means method highly depend on the initialization of the initial cluster center values provided (Anggriawan & Gunawan, 2022). K-Means is a non-hierarchical method that initially selects a number of population components to be used as the initial cluster centers (Astuti & Yuniarti, 2023). At this stage, the cluster centers are randomly chosen from a set of population data. Next, K-Means tests each component in the data population and assigns it to one of the defined cluster centers depending on the minimum distance between the component and each cluster. The cluster center positions are recalculated until all data components are classified into each cluster center, resulting in new cluster center positions. The steps for clustering using the K-Means method are as follows:

1. Choose the number of clusters k.
2. Initialize k cluster centers. This can be done in various ways, but the most common method is random initialization. The cluster centers are given initial values with random numbers.
3. Assign all data/objects to the nearest cluster. The proximity of two objects is determined based on the distance between them. Similarly, the proximity of a data point to a particular cluster is determined by the distance between the data point and the cluster center. In this stage, the distance of each data point to each cluster center must be calculated. The smallest distance between a data point and a particular cluster determines which cluster the data point belongs to. To calculate the distance of all data points to each cluster center, the Euclidean distance theory can be used, formulated as follows:

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2}$$

Square Within-Cluster (SSW)

To determine the cohesion within a cluster, one method is to calculate the value of the Sum of Square Within-Cluster (SSW) using the following formula:

$$SSW\ i = \frac{1}{mi} \sum mi/j = i\ d(Xj, Cj)$$

Where:
mi = the number of data points in cluster i
ci = the centroid of cluster i
d (Xj, Cj) = the distance of each data point to centroid i, calculated using the Euclidean distance.

Sum Of Square Between-Cluster (SSB)

The calculation of Sum of Square Between-Cluster (SSB) aims to determine the separation or distance between clusters using the following formula:

$SSBij = d\ (Xi,Xj)$

Where:

$d\ (Xi,\ Xj)$ = the distance between data point i and data point j in different clusters.

Ratio

The calculation of the ratio (Ri,j) aims to determine the comparative value between cluster i and cluster j to calculate the ratio value for each cluster. The indices i and j represent the number of clusters. If there are 4 clusters, then there are 4 indices, namely i,j,k and l. The ratio value is determined using the following formula:

$$Rij\ldots\ldots,n = \frac{SSWi+SSWj+\cdots+SSWn}{SSBi,j+\cdots+SSBni,j}$$

Where:

$SSWi$ = Sum of Square Within-Cluster at centroid i

$SSBi,j$ = Sum of Square Between Cluster data i and j in different clusters.

Davies Bouldin Index (DBI)

The method for calculating the Davies Bouldin Index (DBI) involves first computing the ratio values, which represent the relationships between Sum of Square Within-Cluster (SSW) and Between-Cluster (SSB) sums. Subsequently, DBI is computed using the following formula:

$$DBI = \frac{1}{k}\sum\frac{k}{i=1}\ max\ i{\neq}j\ (Ri,j,\ \ldots\ k)$$

Where Ri,j represents the ratios calculated from the SSW and SSB values. A lower DBI indicates better clustering quality, with k representing the number of clusters. From the calculation of Davies Bouldin Index (DBI), it can be concluded that the smaller the DBI value obtained (non-negative ≥ 0), the better the cluster quality.

RESEARCH METHODS

This research will go through several processes. The sequence of processes in this research can be seen in the following diagram:

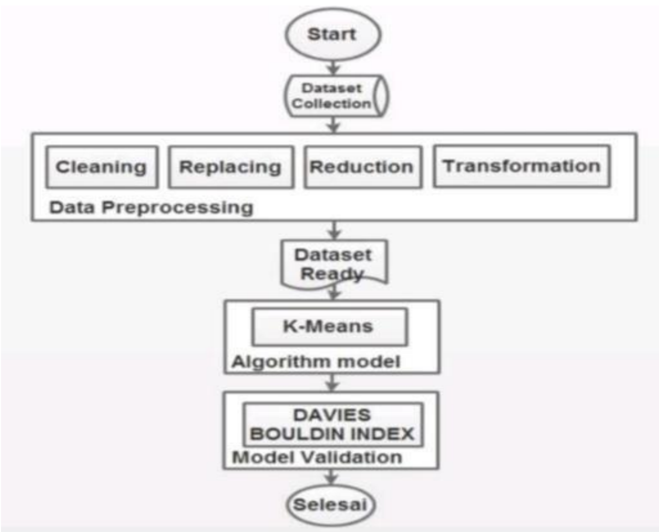

Figure 1. Research Methodology

The initial stage of the research begins with selecting the dataset to be used. In this study, the dataset used is car sales data. The data collected comprises car sales data from Gaikindo (Association of Indonesian Automotive Industries). The data obtained includes car sales results in Indonesia from 2015 to 2019. Data collection is based on the requirements of this research. With this data, it is expected to obtain relevant, accurate, and reliable information. The data collection is carried out by retrieving car sales data from the Gaikindo website, which will then be processed using the K-Means algorithm.

RESULTS AND DISCUSSION

In this study, the dataset consists of a total of 900 monthly car sales records over 5 years. The data is gathered based on sales summaries obtained from the official Gaikindo website for research purposes, which will later be compiled into an Excel file. The data collected includes car sales from the year 2015 to 2019, categorized by month, brand, and sales figures. With a total of 900 data points, the distribution of the data is as follows:

Table 1. Dataset Results

NO.	BULAN	BRAND	PENJUALAN
1	Jan-15	Mitsubishi	10.257
2	Jan-15	Datsun	2.015
3	Jan-15	Nissan	1.985
4	Jan-15	Hino	1.948
5	Jan-15	Isuzu	1.372
6	Jan-15	Suzuki	11.687
7	Jan-15	Chevrolet	644
8	Jan-15	Toyota	22.555
9	Jan-15	Daihatsu	14.701
10	Jan-15	BMW	180
11	Jan-15	Hyundai	167
12	Jan-15	Lexus	8
13	Jan-15	Audi	19
14	Jan-15	Honda	11.354
15	Jan-15	Mazda	711
16	Jan-15	Mitsubishi	8.968
17	Jan-15	Datsun	2.067
18	Jan-15	Nissan	3.007
...	....	....	....
898	Des-19	Audi	4
899	Des-19	Honda	14.363
900	Des-19	Mazda	368

Next, the data in table 2 will undergo transformations to make it easier to process. Since some attributes of the data used are non-numeric, it is necessary to convert this data into numeric format.

Table 2. Data Transformation Table

NO.	BRAND	PENJUALAN
1	1	10.257
2	2	2.015

3	3	1.985
4	4	1.948
5	5	1.372
6	6	11.687
7	7	644
8	8	22.555
9	9	14.701
10	10	180
11	11	167
12	12	8
13	13	19
14	14	11.354
15	15	711
16	16	8.968
17	17	2.067
18	18	3.007
...	....	....
898	898	4
899	899	14.363

Table 3. Initial Centroids Iteration 1

	BRAND	**PENJUALAN**
C1	3	1.985
C2	4	1.948
C3	5	1.372

Calculate the distance between data points and Centroid using Euclidean Distance with the following formula:

$$D(i,j)= \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}$$

The calculation of the first distance between the first data point and the first cluster centroid is:

a. d1,c1 (data 1, centroid 1)

$$= \sqrt{(1-3)^2 + (10.257 - 1.985)^2}$$
$$= 8272$$

b. d1,c2 (data 1, centroid 2)

$$= \sqrt{(1-4)^2 + (10.257 - 1.948)^2}$$
$$= 8309{,}001$$

c. d1, c3 (data 1, centroid 3)

$$= \sqrt{(1-5)^2 + (10.257 - 1.372)^2}$$
$$= 8888{,}001$$

Table 4. Calculation Results of Distance from Cluster Centers Iteration 1

C1	C2	C3	Jarak Terpendek
8272	8309	8885	8272
30	67	643	30
0	37	613	0
37	0	576	0
613	576	0	0
9702	9739	10315	9702
1341	1304	728	728
20570	20607	21183	20570
12716	12753	13329	12716
1805	1768	1192	1192
1818	1781	1205	1205
1977	1940	1364	1364
1966	1929	1353	1353
9369	9406	9982	9369
1274,1	1237	661,1	661,1
6983	7020	7596	6983
82	119	695	82
1022	1059	1635	1022
...	...	...	...
1981	1944	1368,	1368
12378	12415	12991	12378
1617	1580	1004	1004

Group the data according to their clusters by assigning each data point to the cluster with the shortest distance. For example, in Table 4, it can be observed that the data points have a smaller distance to centroid 1 compared to centroids 2 and 3. An example of grouping data by closest distance can be seen in Table below:

Table 5. Data Distance and Centroid Grouping Iteration 1

No.	Brand
1	1
2	1
3	1
4	2
5	3
6	1
7	3
8	1
9	1
10	3
11	3
12	3
13	3
14	1
15	3
16	1
17	1
18	1
...	....
898	3
899	1
900	3

Proceeding back to step 2 involves using the new centroids from the first iteration, which are calculated from the average values of each cluster group. The new centroids are obtained by summing all attribute values of data points assigned to a centroid and dividing by the number of data points, and this applies to all centroid attributes. For example, for the 'brand' attribute in the first centroid:

Many data in the first cluster (c1)

$$= \frac{2591}{384} = 6{,}747396$$

Table 6. New Centroids Data Iteration 2

	Brand	Penjualan
C1	6,7	12992,2
C2	4,2	1824,7
C3	9,3	408,1

Table 7. New Centroids Data Iteration 3

	Brand	Penjualan
C1	7,7	16363,4
C2	3,9	2364,1
C3	10,1	275,2

Table 8. New Centroids Data Iteration 4

	Brand	Penjualan
C1	8,3	17705,7
C2	4	3678,6
C3	9,8	321,5

Table 9. New Centroids Data Iteration 5

	Brand	Penjualan
C1	8,8	18900,5
C2	4,3	5830,5
C3	8,9	523,6

Table 10. New Centroids Data Iteration 6

	Brand	Penjualan
C1	8,9	20014,7
C2	5,4	8945,6
C3	8,3	761,5

Table 11. New Centroids Data Iteration 7

	Brand	Penjualan
C1	8,6	22317,4
C2	6,8	10723,7
C3	8,2	823

Table 12. New Centroids Data Iteration 8

	Brand	Penjualan
C1	8,3	25802
C2	7,3	11852,7
C3	8,2	838,3

Table 13. New Centroids Data Iteration 9

	Brand	Penjualan
C1	8,5	28420,6
C2	7,4	12356,5
C3	8,2	838,3

Table 14. New Centroids Data Iteration 10

	Brand	Penjualan
C1	8,1	29293,4
C2	7,5	12537,2
C3	8,2	838,3

Table 15. New Centroids Data Iteration 11

	Brand	Penjualan
C1	8	29437,2
C2	7,5	12571,2
C3	8,2	838,3

The next process involves iterating with the new centroids to achieve convergence, where the average values stabilize. In this study, the iteration process with the new centroids stops at the 11th iteration. The final centroids that do not change can be observed in Table below:

Table 16. Latest Iteration Centroid Data

	Brand	Penjualan
C1	8	29437,2
C2	7,5	12571,2
C3	8,2	838,3

The following are the results from the last iteration:

Table 17. Results of the 11th Iteration (Final)

No.	Bulan	Brand	Penjualan	C1	C2	C3	Jarak Terpendek	k
1	Jan-15	1	10.257	19036,4	2280,2	9418,7	2280,2	2
2	Jan-15	2	2.015	27278,4	10522,2	1176,7	1176,7	3
3	Jan-15	3	1.985	27308,4	10552,2	1146,7	1146,7	3
4	Jan-15	4	1.948	27345,4	10589,2	1109,7	1109,7	3
5	Jan-15	5	1.372	27921,4	11165,2	533,7	533,7	3
6	Jan-15	6	11.687	17606,4	850,2	10848,7	850,2	2
7	Jan-15	7	644	28649,4	11893,2	194,3	194,3	3
8	Jan-15	8	22.555	6738,4	10017,8	21716,7	6738,4	1
9	Jan-15	9	14.701	14592,4	2163,8	13862,7	2163,8	2
10	Jan-15	10	180	29113,4	12357,2	658,3	658,3	3
11	Jan-15	11	167	29126,4	12370,2	671,3	671,3	3
12	Jan-15	12	8	29285,4	12529,2	830,3	830,3	3
13	Jan-15	13	19	29274,4	12518,2	819,3	819,3	3
14	Jan-15	14	11.354	17939,4	1183,2	10515,7	1183,2	2
15	Jan-15	15	711	28582,4	11826,2	127,5	127,5	3
16	Feb-15	16	8.968	20325,4	3569,2	8129,7	3569,2	2
17	Feb-15	17	2.067	27226,4	10470,2	1228,7	1228,7	3
18	Feb-15	18	3.007	26286,4	9530,2	2168,7	2168,7	3
---	......	......	....	---------	-------	-------	------	--
898	Des-19	898	4	29289,4	12533,2	834,3	834,3	3
899	Des-19	899	14.363	14930,4	1825,8	13524,7	1825,8	2
900	Des-19	900	368	28925,4	12169,2	470,4	470,4	3

To obtain the Davies Bouldin Index value, first calculate the Sum of square within-cluster, Sum of square between-cluster, and Ratio. To calculate the Davies Bouldin Index, the data used is the data that has formed in the final clustering and the centroid of the last cluster as shown in previous table.

$$SSW_{0} = \frac{6738,4 + 5656,4 + 858,4 + \cdots + 4126,6}{61}$$

$$= 6012{,}2$$

$$SSW_1 = \frac{2280{,}2 + 850{,}2 + 2163{,}8 + \cdots + 1825{,}8}{235}$$

$$= 2945{,}1$$

$$SSW_2 = \frac{1176{,}7 + 1146{,}7 + 1109{,}7 + \cdots + 470{,}4}{604}$$

$$= 10532{,}9$$

After determining the SSW value, the next step is to calculate the Sum of square between-cluster (SSB). To calculate the SSB value, the final centroid from the last iteration is needed. Below is the final centroid obtained from the last iteration.

Table 18. Final Centroids from the Clustering Process

	Brand	Penjualan
C1	8	29437,2
C2	7,5	12571,2
C3	8,2	838,3

The SSB calculation is performed as follows

$$SSB_{1,2} = \sqrt{(8 - 7{,}5)^2 + (29437{,}2 - 12571{,}2)^2}$$

$$= 16866$$

$$SSB_{1,3} = \sqrt{(8 - 8{,}2)^2 + (29437{,}2 - 838{,}3)^2}$$

$$= 28598{,}9$$

$$SSB_{2,3} = \sqrt{(7{,}5 - 8{,}2)^2 + (12571{,}2 - 838{,}3)^2}$$

$$= 11732{,}9$$

After the SSW and SSB values have been calculated and the results obtained, the next step is to find the inter-cluster ratio with the following calculation:

$$R_0 = \frac{6012{,}2}{16866 + 28598{,}9 + 11732{,}9}$$

$$= \frac{6012{,}2}{57197{,}8} = 0{,}105$$

$$R_1 = \frac{2945{,}1}{16866 + 28598{,}9 + 11732{,}9}$$

$$= \frac{2945,1}{57197,8} = 0,051$$

$$R_1 = \frac{10532,9}{16866 + 28598,9 + 11732,9}$$

$$= \frac{10532,9}{57197,8} = 0,184$$

After determining the inter-cluster ratio, calculate the DBI value as follows:

$$DBI = \frac{R0+R1+R2}{K}$$

$$= \frac{0,105 + 0,051 + 0,184}{3}$$

= 0,341

In this Implementation and Testing, the researcher will use the software RapidMiner Studio Version 9.7.002. By testing data using this software, we will compare the results of manual data processing with the results of data processing using software. To import data into the RapidMiner Studio v.9.7.002 application, there are 3 stages. The steps are as follows:

1. Locate the file

The first stage is to locate the file that was previously created in .xlsx or .xls format, select it, and then save it. In this test, the data to be tested is saved under the name Data Penjualan Mobil Gaikindo.xlsx, then select it.

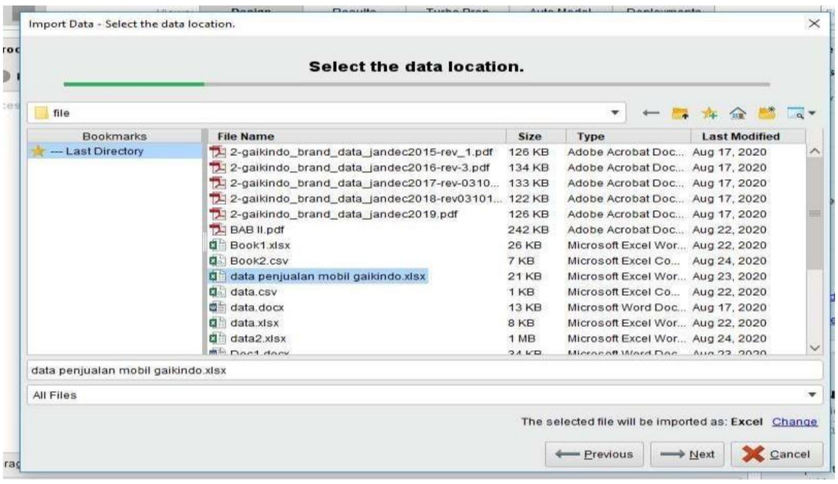

Figure 2. File location search display

2. Select Sheet

Then select next then proceed to the next stage selecting the Sheet which contains the data source in Microsoft Excel that is used, as shown in Figure below.

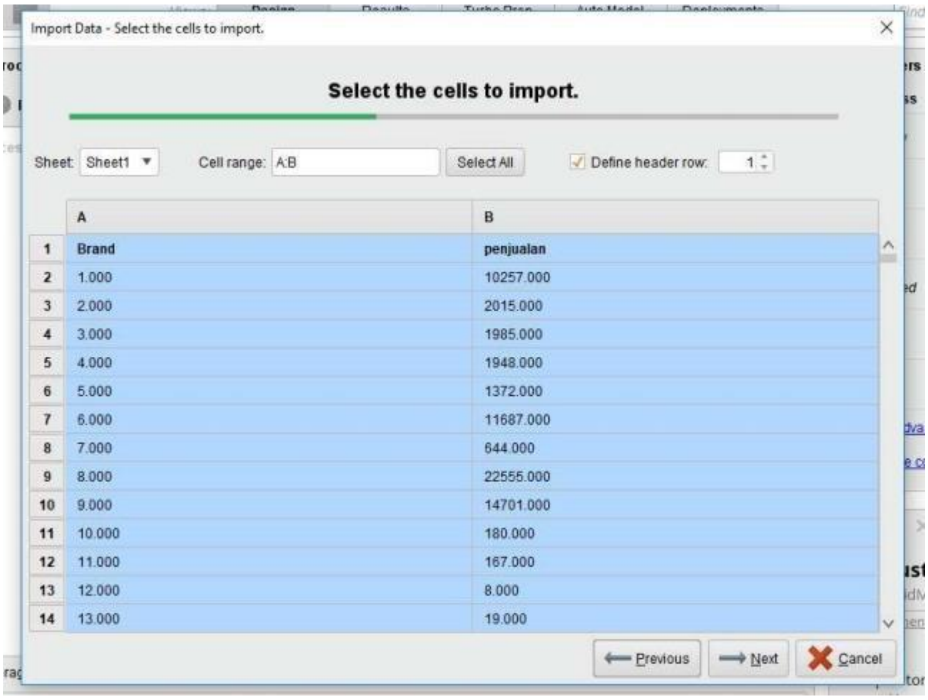

Figure 3. Sheet selection display

3. Annotation

This stage is the annotation stage. If our data does not have attribute names, there is no need to do anything at this stage. Then click the Next button.

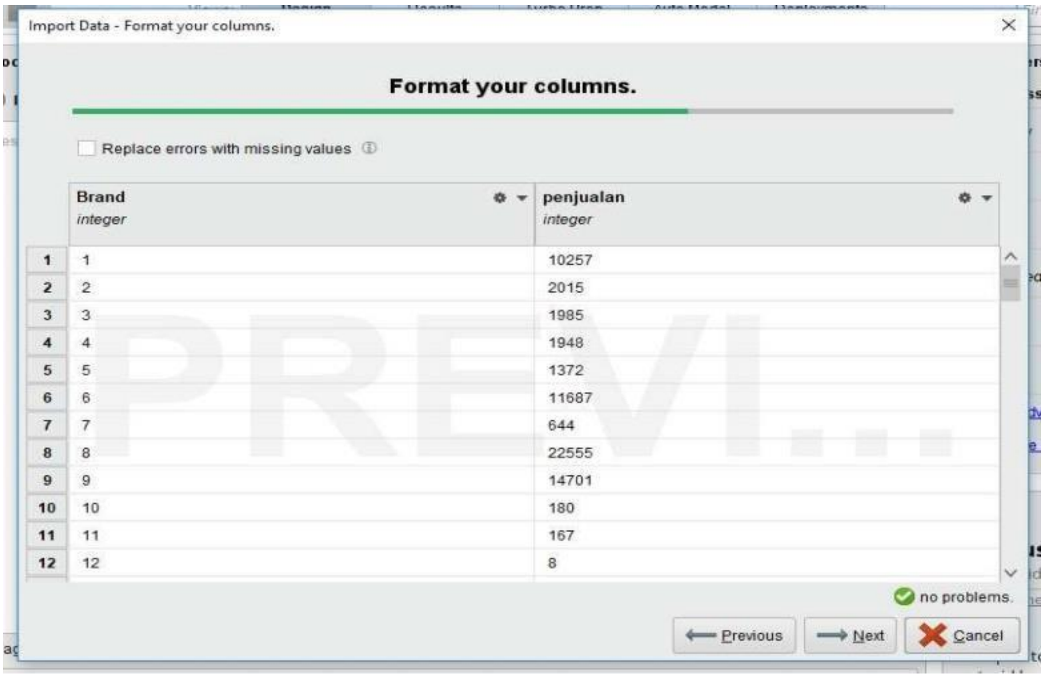

Figure 4. Annotation display

The data import stage is completed by selecting Finish, then in the Main Process you will see a new operator which already contains the Gaikindo.xlsx Car Sales Data file which was imported directly from the excel file. The data is ready for testing.

The next step is to add the K-Means operator by directly typing "K-Means" in the Search For Operator column, click and hold then drag it to Process.

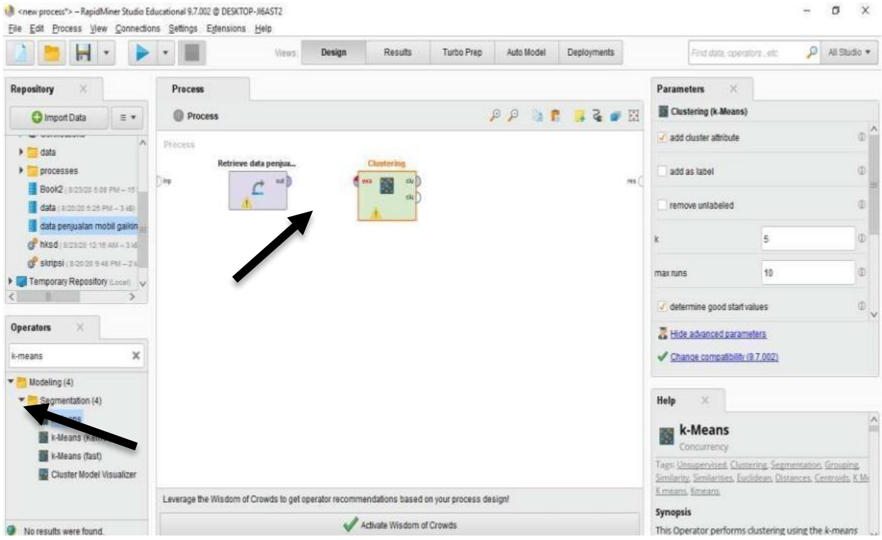

Figure 5. Adding the K-Means Operator

Next, configure the K-Means algorithm, which is set in the K-Means Clustering Parameter menu, as shown in Figure below.

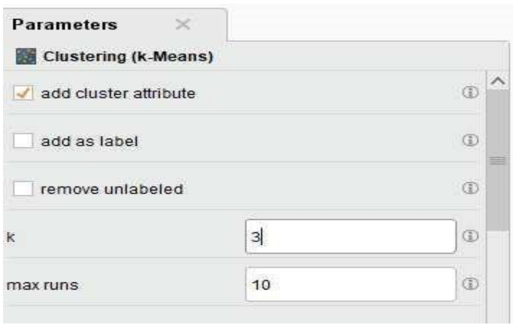

Figure 6. Determining the Number of Clusters

In Figure above, the value of k is set, where k is the number used to determine the number of clusters to be formed. Here, the number of clusters to be formed is 3, corresponding to sales levels: less popular, popular, and most popular. The next step is to add the Cluster Distance Performance operator by typing "Cluster Distance Performance" directly into the Search for Operator column, clicking and holding, then dragging it into the Process.

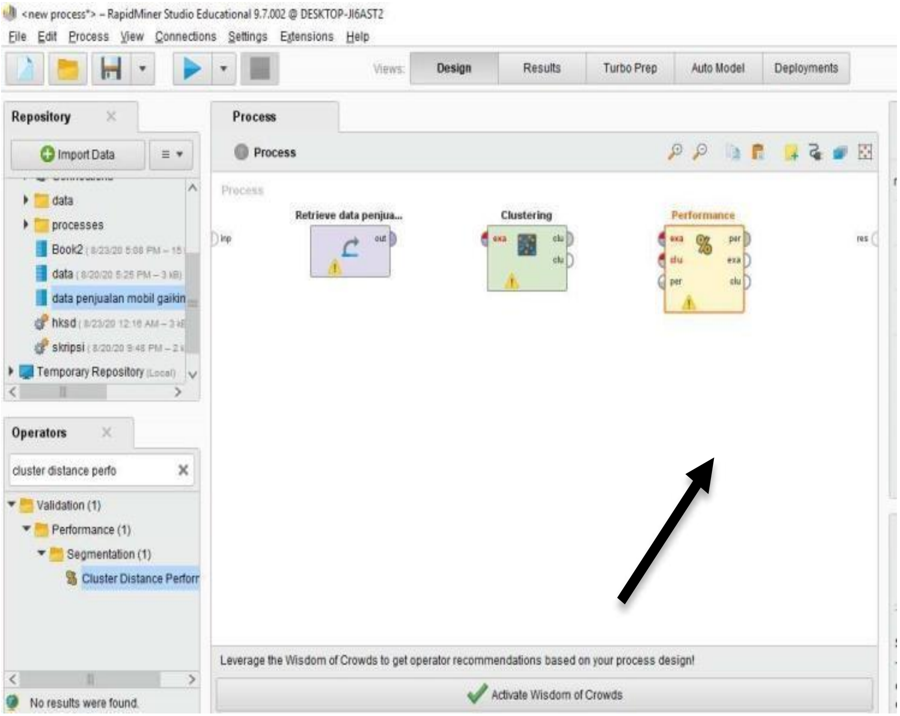

Figure 7. Adding OperaCtolurster Distance Performance

Connect the data with Clustering K-Means and Cluster Distance Performance to determine the output towards the result. Finally, click the Play button as shown below.

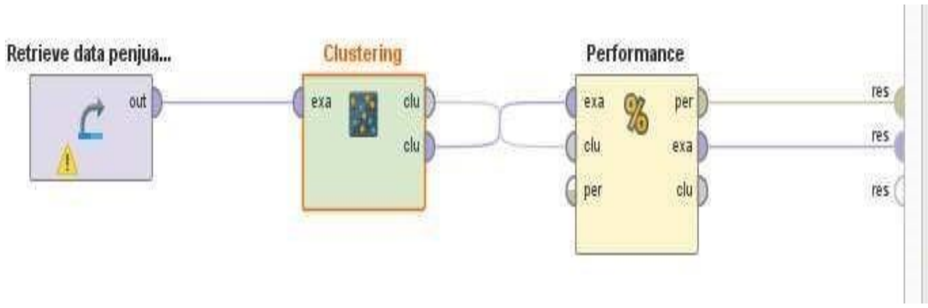

Figure 8. Display for the K-Means Clustering Process

There are 3 processes that will be carried out at this stage, namely:

1. Retrieve test data

   This stage carries out the operation of inputting the dataset in the form of a file with the extension.xls car sales data.

2. Clustering

   This stage carries out clustering operations as an algorithm carried out in this research.

3. Performance

   This stage carries out a search operation for the Davies Bouldin Index.

After running and undergoing the 3 stages above, there are several output results in testing by rapidminer, namely as follows:

ExampleSet

In ExampleSet you can see several cluster results displays, namely Data View. Data View is a display of the results of the data cluster as a whole according to the data that has been input.

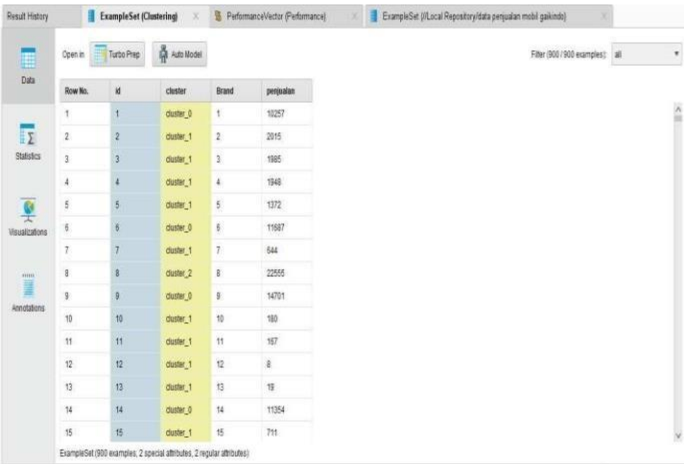

Figure 9. ExampleSet display

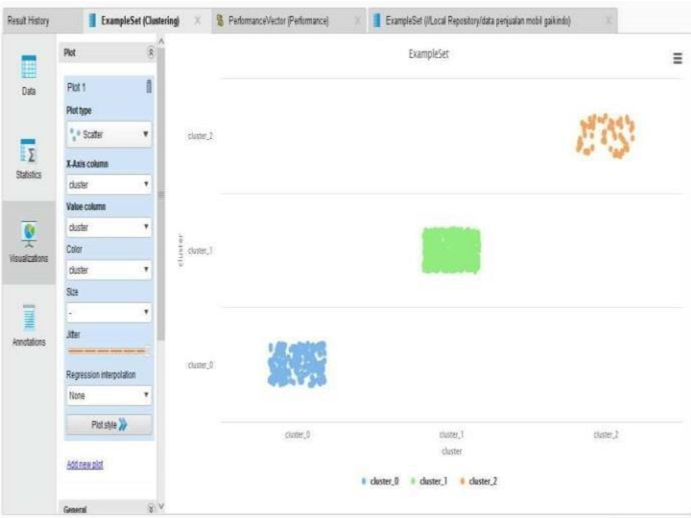

Figure 10. Scatter display

At this stage, the results of data grouping are displayed in the form of a dot graph in three colors for each cluster. The blue color represents cluster 0 with 235 members, green cluster 2 with 604 members, red cluster 1 with 61 members from the 900 datasets that have been tested.

Cluster Model

```
Cluster 0: 235 items
Cluster 1: 604 items
Cluster 2: 61 items
Total number of items: 900
```

Figure 11. Text View Cluster Model display

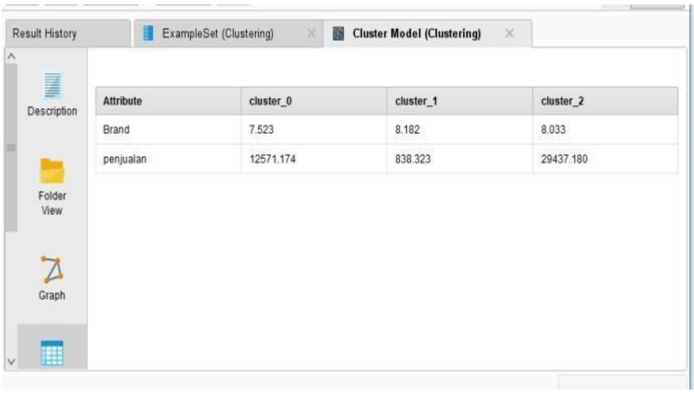

Figure 12. Centroid Table Display

At this stage, the central point value for each cluster is displayed. This value will be a reference for calculations in each dataset by measuring the value with each cluster center point.

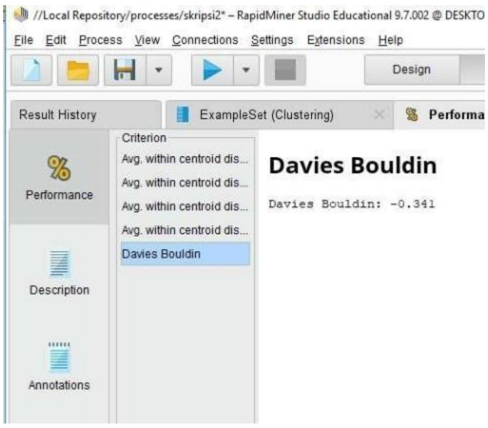

Figure 13. View of the Davies Bouldin Index

The smaller the Davies Bouldin Index value, the better the cluster obtained from grouping using the clustering method. The calculation results using the k-means algorithm show a value of 0.341.

After grouping and testing manually and also testing using RapidMiner, it can be concluded that of the three clusters, namely Cluster 0 (In demand) with 235 members, Cluster 1 (Less in demand) with 604 members, Cluster 2 (Very in demand) with the number of members is 61. Based on the results of the analysis, the presentation for each cluster is 26% for Cluster 0, 67% for Cluster 1, and 7% of the data for Cluster 2.

Table 19. Clustering results manually and using rapidminer

Cluster 0 (C0)	235	26%	Laris
Cluster 1 (C1)	604	67%	Kurang Laris
Cluster 2 (C2)	61	7%	Sangat Laris
Jumlah	900	100%	

CONCLUSION

From the research conducted by the researcher, the following conclusions are drawn:

1. Clustering method using the K-Means algorithm can be used to group car sales data based on sales volumes over 5 years from 900 data points into Cluster 0 (Popular), Cluster 1 (Less Popular), and Cluster 2 (Very Popular). Therefore, companies whose car sales fall into Cluster 1 can strategize to increase their sales.
2. This clustering method can assist automotive companies in comparing their car sales levels with others. Based on testing results, both manual and using RapidMiner v9.7.002 software, identical and accurate results were obtained: Cluster 0 with 235 members, accounting for 26% (Popular); Cluster 1 with 604 members, accounting for 67% (Less Popular); and Cluster 2 with 61 members, accounting for 7% (Very Popular). The Davies Bouldin Index (DBI) calculated using the K-Means algorithm in this study showed a value of 0.341.
3. Further research could expand using similar data and methods, employing combinations or alternative approaches to achieve more robust research outcomes.

REFERENCES

Anggriawan, I., & Gunawan, W. (2022). Implementation of Data Mining Using K-Means Algorithm for Bicycle Sales Prediction. ILKOM Jurnal Ilmiah, 14(3), 284–293. https://doi.org/10.33096/ilkom.v14i3.1291.284-293

Annur, H. (2019). Penerapan data mining menentukan strategi penjualan variasi mobil menggunakan metode K-Means clustering. *J. Inform. Upgris, 5*(1). https://doi.org/10.26877/jiu.v5i1.3091

Astuti, J., & Yuniarti, T. (2023). Data Mining Modeling in Clustering Car Products Sales Data in the Automotive Industry in Indonesia. Jurnal Manajemen Industri Dan Logistik, 7(2), 261–281. https://doi.org/10.30988/jmil.v7i2.1258

Butsianto, S., & Saepudin, N. (2020). Penerapan data mining terhadap minat siswa dalam mata pelajaran matematika dengan metode K-Means. *J. Nas. Komputasi dan Teknol. Inf., 3*(1), 51–59. https://doi.org/10.32672/jnkti.v3i1.2008

Fatmawati, K., & Windarto, A. P. (2018). Data mining: Penerapan Rapidminer dengan K-Means cluster pada daerah terjangkit demam berdarah dengue (DBD) berdasarkan provinsi. *Comput. Eng. Sci. Syst. J., 3*(2), 173. https://doi.org/10.24114/cess.v3i2.9661

Ginting, B., & Riandari, F. (2020). Implementasi metode K-Means clustering dalam pengelompokan bibit tanaman kopi arabika. *J. Nas. Komputasi dan Teknol. Inf., 3*(2).

Govindasamy, K., & Velmurugan, T. (2017). A Study on Classi fi cation and Clustering Data Mining Algorithms based on Students Academic Performance Prediction. International Journal of Control Theory and Applications, 10(23), 147–160.

Handoko, K. (2016). Penerapan data mining dalam meningkatkan mutu pembelajaran pada instansi perguruan tinggi menggunakan metode K-Means clustering (Studi kasus di program studi TKJ Akademi Komunitas Solok Selatan). *J. Teknol. dan Sist. Inf., 02*(03), 31–40. Retrieved from http://teknosi.fti.unand.id/index.php/teknosi/article/view/70

Hutasoit, M. N., Fa'rifah, R. Y., & Andreswari, R. (2023). Application of Data Mining For Clustering Car Sales Using The K-Means Clustering Algorithm. International Journal of Information System & Technology Akreditasi, 7(2), 118–124.

López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. Psicothema, 33(3), 456–465. https://doi.org/10.7334/psicothema2021.62

Maulida, L. (2018). Penerapan data mining dalam mengelompokkan kunjungan wisatawan ke objek wisata unggulan di Prov. DKI Jakarta dengan K-Means. *JISKA (Jurnal Inform. Sunan Kalijaga), 2*(3), 167. https://doi.org/10.14421/jiska.2018.23-06

Metisen, B. M., & Sari, H. L. (2015). Analisis clustering menggunakan metode K-Means dalam pengelompokkan penjualan produk pada Swalayan Fadhila. *J. Media Infotama, 11*(2), 110–118.

Rezaeinejad, I. (2021). Automotive Industry and Its Place in the Economy: Case Study Iran Auto

Industry. Management, 5(3), 23–32. https://www.researchgate.net/publication/354677669

Rony, S. (2016). Penerapan data mining menggunakan algoritma K-Means clustering untuk menentukan strategi promosi mahasiswa baru (Studi kasus: Politeknik LP3I Jakarta). *J. Lentera ICT, 3*(1), 76–92.

Setiyani, L., & dkk. (2020). Analisis prediksi kelulusan mahasiswa tepat waktu menggunakan metode data mining Naïve Bayes: Systematic review. *Fakt. Exacta, 13*(1), 38–47. https://doi.org/10.30998/faktorexacta.v13i1.5548

Silalahi, R. E., & dkk. (2019). Penerapan data mining dalam menentukan penerima bantuan sosial dengan algoritma C4.5 pada kantor kecamatan Siantar Selatan. *J. Sist. Inf. Ilmu Komput. Prima (JUSIKOM PRIMA), 1*, 452–460.

Tamba, S. P., & dkk. (2019). Penerapan data mining untuk menentukan penjualan sparepart Toyota dengan metode K-Means clustering. *J. Sist. Inf. Ilmu Komput. Prima (JUSIKOM PRIMA), 2*(2), 67–72.