e-ISSN: 3030-802X

GROUPING PRODUCTS IN SUPERMARKETS USING THE K-MEANS ALGORITHM

Mohammad Aldinugroho Abdullah
Universitas Nasional, Jakarta, Indonesia
Email : nugrohoaldi48@gmail.com

Rima Tamara Aldisa*
Universitas Nasional, Jakarta, Indonesia
Correspondence author Email: rimatamaraa@gmail.com

Abstract

Supermarket, a shop that provides various products for use, especially for daily life, including food products, drinks, kitchen utensils, clothing, electronic equipment and others. It is not surprising that many mothers now choose to shop for daily necessities at supermarkets rather than the nearest stall. With self-service, it can make it easier for us consumers to buy different products in one place. So there is no need to change shops to buy other items. Of course, products have different levels of popularity, not only because of taste but also because of price. The number of products provided by supermarkets is relatively large and if you look at the level of popularity, it is difficult to determine, so data mining is needed. The data mining used is clustering. After implementing and using the K-Means algorithm in clustering (grouping) supermarket products, there are two centroids used (C1 for Not Selling Products and C2 for Best Selling Products). The initial centroid value is determined randomly, while the subsequent centroids are adjusted according to the results of calculating the closest distance (maximum distance). The final result obtained is that the best-selling group consists of 12 products, namely products with serial numbers 1, 4, 5, 6, 7, 8, 9, 11, 14, 15, 16 and 17. Meanwhile, the product group does not There are 6 best-selling products, namely products with serial numbers 2, 3, 10, 12, 13 and 18.

Keywords: Data mining, Clustering, K-Means, Self-Service Products

INTRODUCTION

Supermarkets are shops that provide various products for use, especially for daily life, including food products, drinks, kitchen utensils, clothing, electronic equipment and others. It is not surprising that many mothers now choose to shop for daily necessities at supermarkets rather than the nearest stall. With self-service, it can make it easier for us consumers to buy different products in one place. So there is no need to change shops to buy other items. Of course, products have different levels of popularity, not only because of taste but also because of price. The number of products provided by supermarkets is relatively large and if you look at the level of popularity, it is difficult to determine, so data mining is needed.

Data mining is a process of artificial intelligence, machine learning and statistics to analyze and identify large amounts of information (data) [1]. The data mining groupings are association, description, estimation, prediction, classification and clustering. This research focuses on the clustering method because it is in accordance with the aim of this research, namely grouping.

Clustering is a data mining method, used to group data, grouping is done based on the similarity of the data [2][3][4]. The k-Means algorithm was used for this research.

The initial step is to determine how many clusters are formed, then determine the initial cluster by selecting the data randomly, k-means is applied repeatedly or known as iteration. The iteration process will not stop if the results of the minimum distance calculation change, and vice versa, if the results remain the same as the previous iteration then the iteration process stops [5].

Previous research conducted by Riyani Wulan Sari and Dedy Hartama in 2018, this research discussed how to develop the potential of objects in each province of Indonesia so that they can make foreign tourists interested in visiting these places. The method used for grouping is the K-Means algorithm [6]. Further research was conducted in 2019 by Leonardo Purba, et al. With data mining research in grouping provinces affected by AIDS, the algorithm used in the clusteringis K-Medoids with research results namely that there are 5 provinces grouped into the lowest value (cluster 1) and there are 28 provinces grouped into the high value (cluster 2) [7]. Further research by Mawaddah Anjelita, et al. in 2019, the results were that Cluster 1 had 4 provinces experiencing high levels of water pollution and Cluster 2 had 30 provinces with low levels of water pollution. This research uses the K-Miens algorithm to group provinces that experience water pollution [8]. The latest research was conducted in 2019 by researcher Dewinta Marthadinata Sinaga, et al. with research that discusses how to group consumer price indexes using the K-Means algorithm. The results of this research show that there are 14 cities grouped into Cluster 1, then there are 29 cities grouped into Cluster 2 and 23 cities in Cluster 3 [9].

RESEARCH METHODOLOGY

The process which takes place at the time of completiona research from start to finish, for research stages so that the completion of this research is carried out systematically and structured. The following is Figure 1, the stages of this research:

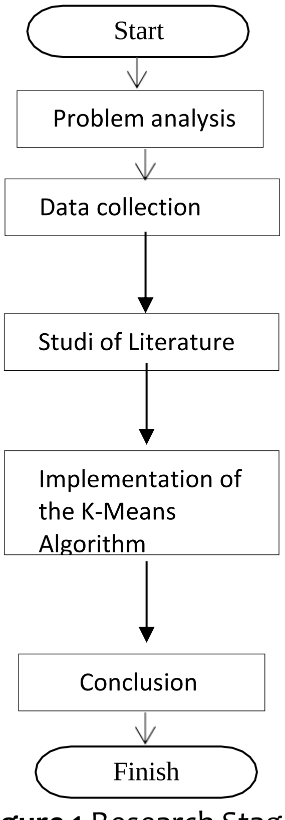

Figure 1. Research Stages

1. Analyze problems that occur in supermarkets and create solutions to these problems
2. Collect product data in supermarkets at the research location and adapt it to the problems experienced.
3. The aim of the literary study is thatResearchers can understand the problem solving process based on existing research and relate it to the problems in this research.
4. The K-Means algorithm is often used to group products in supermarkets.
5. Make a conclusion after carrying out the problem analysis process and applying the K-Means algorithm to produce research results.

Data Mining

The process of analyzing several machine learning or computer learning techniques so that they can extract knowledge automatically is an understanding of data mining. Data mining is very useful for the future because in data mining a pattern will be formed that is used to make decisions. The data that is processed and created patterns certainly has a large database so data mining is really needed [10]. Data mining is divided into several groups and each problem is adapted to its respective solution group. The groupings used in data mining are association, description, estimation, prediction, classification and clustering [11] [12]:

Clustering

Datamining is used for grouping, forming several objects and overall observations based on the similarities of these objects. Clustering is not only used for grouping based on similarity, but can also be used for groups that have no similarities. The way clustering works is different from other data mining groups where clustering does not carry out prediction, calcification and estimation actions, clustering is only processed based on the level of similarity of clusters to each other [13].

K-Means

It is a data analysis clustering algorithm to group research objects based on similarities between one characteristic and another. Grouping data in K-Means is carried out using a partition system by carrying out unsupervised modeling [14] [15].

Following are the steps for implementing the K-Means algorithm[10] [16] :

1. Determine (K) the total clusters that will be created
2. Determine the initial centroid
   $$Ci = \frac{1}{M}\sum_{j=1}^{M} X_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$
3. Calculate the distance between the data and the initial centroid
   $$d_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

   Note:
   d(x,y)= distance of data from x to the center of cluster y Xi= ith data in the nth data attribute Yi= jth data in nth data attribute
4. Group the data into the relevant centroid based on the closest distance (minimum distance) from the clusters that have been calculated.

5. After obtaining the cluster grouping based onclosest distance, then carry out iteration using the new centroid value. The new centroid value is adjusted to the location of the minimum distance cluster then do equation one.
6. If the centroid class grouping results move, then repeat step 3. If the data grouped by centroid class does not move, then the process stops (the iteration is complete).

RESULT AND DISCUSSION

Grouping products in a supermarket based on the product's popularity level by clustering product data using the k-means algorithm. K-means is inseparable from iteration, where the iteration process will not stop if the final result of the shortest distance changes and vice versa, if the result of the shortest distance does not change from the previous iteration, the iteration process will stop. In this solution, the number of centroids or groupings will be determined. In this research, two centroids are needed (Best Selling Products and Not Selling Products).

The data sample used was 18 data with total sales for three months (August, September and October). sample data is shown in table 1.

Table 1. Data Sample

Number.	Item code	Sales Amount		
		August	September	October
1	PB0001	15	24	30
2	PB0002	13	5	21
3	PB0003	8	3	26
4	PB0004	18	33	44
5	PB0005	5	41	28
6	PB0006	9	27	42
7	PB0007	4	19	31
8	PB0008	31	17	36
9	PB0009	27	38	30
10	PB0010	7	25	14
11	PB0011	19	40	29
12	PB0012	3	24	16
13	PB0013	33	15	8
14	PB0014	24	28	30
15	PB0015	8	22	39
16	PB0016	51	36	21
17	PB0017	53	33	19
18	PB0018	30	15	10

Stages of Implementing the 1st Iteration K-Means Algorithm

1. Number of clusters formed K=2 (C1 and C2)
2. Center (centroid) of the initial cluster
   The initial centroid center is used as a grouping or cluster used in this research. The initial centroid is determined randomly, so you are free to use any data. For the initial centroid in table 2.

Table 2. Initial Centroid

No.	Item code	Sales Amount August	 September	 October	Cn
3	PB0003	8	3	26	C1 (Not Selling)
16	PB0016	51	36	21	C2 (Best Selling)

Ci = 1 $\sum M\ X$ .... (1)
$\overline{M}j$=1 j

3. Measure the distance between data, cluster centers with Euclidian distance.
   $d_{Euclidean}\,(X, Y) = \sqrt{\sum_{i=1}^{n}(Xi - Yi)2}$ .... (2)

Data 1:
$C1 = \sqrt{(15 - 8)2 + (24 - 3)2 + (15 - 30)2} = 464$
$C2 = \sqrt{(15 - 51)2 + (24 - 36)2 + (15 - 21)2} = 261$

Data 2:
$C1 = \sqrt{(13 - 8)2 + (5 - 3)2 + (21 - 30)2} = 34$
$C2 = \sqrt{(13 - 51)2 + (5 - 36)2 + (21 - 21)2} = 999$

Data 3:
$C1 = \sqrt{(8 - 8)2 + (3 - 3)2 + (26 - 30)204}$
$C2 = \sqrt{(8 - 51)2 + (3 - 36)2 + (26 - 21)2} = 1157$

Do the steps above until the 18th data. The following is the closest distance (minimum distance) based on the initial centroid in table 3 below.

Table 3. Minimum Distance Iteration 1

No	Item code	C1	C2	Closest distance	Results
1	PB0001	464	261	261	C2
2	PB0002	34	999	34	C1
3	PB0003	0	1157	0	C1
4	PB0004	1234	571	571	C2
5	PB0005	1451	120	120	C2
6	PB0006	833	564	564	C2
7	PB0007	285	436	285	C1
8	PB0008	319	606	319	C1

9	PB0009	1260	109	109	C2
10	PB0010	629	214	214	C2
11	PB0011	1389	112	112	C2
12	PB0012	546	217	217	C2
13	PB0013	493	628	493	C1
14	PB0014	657	172	172	C2
15	PB0015	530	563	530	C1
16	PB0016	1157	0	0	C2
17	PB0017	994	15	15	C2
18	PB0018	422	583	422	C1

4. Data has been grouped into the respective centroid based on the closest distance (minimum distance) from the clusters that have been calculated.
5. After obtaining the cluster grouping based onclosest distance, then carry out iteration using the new centroid value. The new centroid value is adjusted to the location of the minimum distance cluster then do equation one.

   $$C_i = \frac{1}{M}\sum_{j=1}^{M} X_j$$

   $$C_1(Agustus) = \frac{1}{7}(13 + 8 + 4 + 31 + 33 + 8 + 30) = 18.143$$

   $$C_1(September) = \frac{1}{7}(5 + 3 + 19 + 17 + 15 + 22 + 15) = 13{,}714$$

   $$C_1(Oktober) = \frac{1}{7}(21 + 26 + 31 + 36 + 8 + 39 + 10) = 24{,}429$$

   Do the same thing to find the next centroid (C2). The following is a table of the 4 initial centroids that have been calculated
   based on equation 1.

Table 4. Initial Centroid of iteration 2

Centroid	Sales Amount		
	AugustSeptember		October
C1	18,14313,714		24,429
C2	21	31,727	27,545

6. If the centroid class grouping results move, then repeat step 3. If the data grouped by centroid class does not move, then the process stops (iteration is complete).

After calculating, the iteration process stops at the 6th iteration, the results of the 5th and 6th iterations are the same (there is no change in the position of the centroid or the closest distance remains). The following can be seen in table 5, the minimum distance for the 5th iteration and minimum distance for the 6th iteration be seen in the 6th table.

Table 5. Minimum Iteration Distance 5

No	Item code	C1	C2	Closest distance	Results
1	PB0001	243,722	46,090	46,090	C2
2	PB0002	85,056	733,590	85,056	C1
3	PB0003	180,389	775,257	180,389	C1
4	PB0004	1034,722	198,257	198,257	C2
5	PB0005	851,389	137,924	137,924	C2
6	PB0006	744,722	159,590	159,590	C2
7	PB0007	200,722	144,924	144,924	C2
8	PB0008	338,389	216,757	216,757	C2
9	PB0009	739,389	65,257	65,257	C2
10	PB0010	156,056	309,257	156,056	C1
11	PB0011	810,389	99,757	99,757	C2
12	PB0012	125,056	263,424	125,056	C1
13	PB0013	125,722	742,424	125,722	C1
14	PB0014	349,722	7,257	7,257	C2
15	PB0015	504,389	157,090	157,090	C2
16	PB0016	541,056	149,257	149,257	C2
17	PB0017	411,389	167,090	167,090	C2
18	PB0018	85,389	654,090	85,389	C1

Table 6. Minimum Iteration Distance 6

No	Item code	C1	C2	Closest distance	Results
1	PB0001	291,611	43,535	43,535	C2
2	PB0002	119,611	737,701	119,611	C1
3	PB0003	243,278	765,201	243,278	C1
4	PB0004	1137,944	168,201	168,201	C2
5	PB0005	860,944	154,535	154,535	C2
6	PB0006	847,611	129,535	129,535	C2

7	PB0007	261,944	135,701	135,701	C2
8	PB0008	428,278	193,201	193,201	C2
9	PB0009	764,278	74,201	74,201	C2
10	PB0010	122,278	347,535	122,278	C1
11	PB0011	826,944	113,035	113,035	C2
12	PB0012	102,944	295,868	102,944	C1
13	PB0013	78,944	787,201	78,944	C1
14	PB0014	391,278	7,868	7,868	C2
15	PB0015	600,611	130,368	130,368	C2
16	PB0016	524,278	179,035	179,035	C2
17	PB0017	389,611	199,368	199,368	C2
18	PB0018	48,611	693,868	48,611	C1

After the calculation process is complete, products that are selling and not selling can be grouped based on the application of the K-Means algorithm. The following is a table of 7 self-service product groupings:

Table 7. Best Selling Product Group

No	Item code
1	PB0001
4	PB0004
5	PB0005
6	PB0006
7	PB0007
8	PB0008
9	PB0009
11	PB0011
14	PB0014
15	PB0015
16	PB0016
17	PB0017

Table 8. Product Groups Not Selling

No	Item code
2	PB0002
3	PB0003
10	PB0010
12	PB0012
13	PB0013
18	PB0018

Based on table 7, it can be seen that the best-selling product group consists of 12 products and the non-selling products (based on figure 8) there are 6 products so that supermarkets can reduce the number of supplies of these products or can even replace them with other products (new products).

CONCLUSION

The conclusions that can be made after applying the K-Mean algorithm in clustering supermarket products are that there are two centroids used (C1 for non-selling products and C2 for best-selling products). The initial centroid value is determined randomly, while the subsequent centroids are adjusted according to the results of calculating the closest distance (maximum distance). The final result obtained is that the best-selling group consists of 12 products, namely product numbers 1, 4, 5, 6, 7, 8, 9, 11, 14, 15, 16 and 17. Meanwhile, the non-selling product group consists of 6 products, namely products with serial numbers 2, 3, 10, 12, 13 and 18.

REFERENCES

[1] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. Dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019.

[2] S. Al Syahdan and A. Sindar, "Data Mining Penjualan Produk Dengan Metode Apriori Pada Indomaret Galang Kota," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 1, no. 2, 2018, doi: 10.32672/jnkti.v1i2.771.

[3] Z. Nabila, A. R. Isnain, P. Permata, and Z. Abidin, "Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means," *J. Teknol. Dan Sist. Inf.*, vol. 2, no. 2, pp. 100–108, 2021.

[4] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *J. Media Inform. Budidarma*, vol. 4, no. 1, pp. 51–58, 2020.

[5] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, pp. 25–36, 2021.

[6] R. W. Sari and D. Hartama, "Data Mining: Algoritma K-Means Pada Pengelompokkan Wisata Asing ke Indonesia Menurut Provinsi," in *Seminar Nasional Sains dan Teknologi Informasi (SENSASI)*, 2018, vol. 1, no. 1.

[7] L. Purba, S. Saifullah, and R. Dewi, "Pengelompokan Kasus Penyakit Aids Berdasarkan Provinsi Dengan Data Mining K- Medoids Clustering," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, 2019.

[8] M. Anjelita, A. P. Windarto, and D. Hartama, "Pemanfaatan Datamining Pada Pengelompokan Provinsi Terhadap Pencemaran Lingkungan Hidup," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 659–666, 2019, doi: 10.30865/komik.v3i1.1675.

[9] D. M. Sinaga, A. P. Windarto, D. Hartama, and S. Saifullah, "Pengelompokkan Indeks Harga Konsumen Menurut Kota Dengan Datamining Clustering," in *Seminar Nasional Sains dan Teknologi Informasi (SENSASI)*, 2019, vol. 2, no. 1.

[10] Y. Syahra, "Penerapan Data Mining Dalam Pengelompokkan Data Nilai Siswa Untuk Penentuan Jurusan Siswa Pada SMA Tamora Menggunakan Algoritma K-Means Clustering," *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 17, no. 2, p. 228, 2018, doi: 10.53513/jis.v17i2.70.

[11] G. Gunadi and D. I. Sensuse, "Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori Dan Frequent Pattern Growth (Fp-Growth) :," *Telematika*, vol. 4, no. 1, pp. 118–132, 2012.

[12] B. D. Mudzakkir, "Pengelompokan Data Penjualan Produk Pada Pt Advanta Seeds Indonesia Menggunakan Metode K- Means," *J. Mhs. Tek. Inform.*, vol. 2, no. 2, pp. 34–40, 2018.

[13] D. P. Indini, S. R. Siburian, and D. P. Utomo, "Implementasi Algoritma DBSCAN untuk Clustering Seleksi Penentuan Mahasiswa yang Berhak Menerima Beasiswa Yayasan," pp. 325–331, 2022.

[14] F., F. T. Kesuma, and S. P. Tamba, “Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 67–72, 2020, doi: 10.34012/jusikom.v2i2.376.

[15] W. Purba, W. Siawin, and . H., “Implementasi Data Mining Untuk Pengelompokkan Dan Prediksi Karyawan Yang Berpotensi Phk Dengan Algoritma K-Means Clustering,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 85– 90, 2019, doi: 10.34012/jusikom.v2i2.429.

[16] S. A. Rahmah, “KLASTERISASI POLA PENJUALAN PESTISIDA MENGGUNAKAN METODE K-MEANS CLUSTERING (STUDI KASUS DI TOKO JUANDA TANI KECAMATAN HUTABAYU RAJA),” vol. 1, no. 1, pp. 1–5, 2020.