RESEARCH ARTICLE | JUNE 01 2023

Customer clustering using K-means clustering: Supporting customer relationship management system FREE

Silvia Rostianingsih ✉; Alexander Setiawan; Michael Boentarman

AIP Conference Proceedings 2691, 080006 (2023)
https://doi.org/10.1063/5.0116364

Customer Clustering Using K-Means Clustering: Supporting Customer Relationship Management System

Silvia Rostianingsih[a)], Alexander Setiawan[b)], Michael Boentarman[c)]

Informatics Department, Petra Christian University, Surabaya, 60236, Indonesia

[a)]*Corresponding author: silvia@petra.ac.id*
[b)]*alexander@petra.ac.id*
[c)]*michaelboentarman@gmail.com*

Abstract. PT. X is a consumer good company with a variety of products such as consumption salt, packaged cooking oil, soap, sugar, and so on. Currently PT. X has distributors (customers) whose role is to market products directly to consumers. To increase sales from each distributor, PT. X provides special treatment. Special treatment is given manually for certain distributors. This is felt to be less than optimal because the special treatment given is not right on target for each distributor who has the most sold superior product types. Therefore, we need a distributor clustering system using the K-Means method referring to sales history data. The proposed system also has features that can help distributors to place orders independently, confirm payments, submit returns, submit complaints, and exchange reward points for prizes that can be settled by the company. The results of the system have been tested for system acceptance using a questionnaire that was distributed to 5 users. The results of the system acceptance test show the distributor's satisfaction level with the system. Distributors feel the convenience and benefits of the system.

INTRODUCTION

PT. X is a consumer good company with various products such as consumption salt, packaged cooking oil, soap, sugar, and so on. Head office of PT. X is located in Jakarta and has two branch offices in Sidoarjo and Gresik, East Java. As a national company, PT. X has distributors spread across every city throughout Indonesia. The distributor functions as a distributor to sellers in the market, so the role of the distributor has an important role in product sales.

In this study, RFM and K-Means models were applied to cluster distributors based on transaction data. The Frequency pattern analyzes the number of product purchases by distributors, the recency pattern analyzes the length of time for distributors to repurchase, while the monetary pattern analyzes based on the nominal amount of distributor spending on the company. The data analyzed is distributor transaction data in the 2019 period. The results of customer (distributor) clustering analysis using the K-Means method are used as a reference for giving special rewards and discounts for distributors. Rewards are calculated based on the total price of orders made by the distributor. Reward points can be exchanged for various products or vouchers that have been provided by the company. This will support the Customer Relationship Management (CRM). Customer churn analysis supports customer-driven product development with keeping customers engaged and loyal over time [1]. K-Means is chosen since it can analyze large amounts of data and can determine the number of clusters needed and attempt to minimize the total squared error [2]. Though K-Medoids had a better execution time and reduce noise and outlier of data [3] [4]. Thus RFM method helps the company in knowing customer loyalty [5] [6].

STUDY LITERATURE

K-Means clustering is "one of the data methods" non-hierarchical clustering that attempts to partition existing data into one or more forms cluster/group [4]. This method partitions data into clusters/groups so that data that have the

The 6th Engineering Science and Technology International Conference (ESTIC 2021)
AIP Conf. Proc. 2691, 080006-1–080006-7; https://doi.org/10.1063/5.0116364
Published by AIP Publishing. 978-0-7354-4485-0/$30.00

same characteristics are grouped into the same cluster and data that has different characteristics are grouped into other groups.

The basic K-Means algorithm is as follows [2]:

1. Determine the value of k as the number of clusters you want to form.
2. Select the initial K centroid (cluster center point) randomly. In determining the n centers of the initial cluster, a random number is selected which represents the sequence of data inputs. The initial center of the cluster is obtained from the data itself, not by determining a new point, namely by randomizing the initial center of the data.
3. Calculate the distance of each data to each centroid. To measure the distance between data with the center of the cluster used Euclidean distance. Data distance calculation algorithm with the cluster center. The steps are:
   - Take the data value and the cluster center value.
   - Calculate Euclidean distance data with each cluster center. Euclidean distance is the distance obtained from the calculation between all N data with K centroid where will obtain the level of proximity to the closest class to the data population. This Euclidean distance is useful for marking the existence of similarities between each cluster with the minimum distance and has a higher equation. Equality from the Euclidean matrix between the point and the point of the shortest distance as in equation 1.

$$d(x,y) = |x-y| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{1}$$

4. Each data selects the nearest centroid. Distance calculation results will be carried out a comparison and selected the closest distance between the data and the center of the cluster, this distance indicates that the data are in one group with the closest cluster center.
5. Determine the position of the new centroids by calculating the average value of the data located on the same centroid. The new cluster center is used to perform the next iteration if the results obtained have not converged. Process the iteration will stop if it has met the maximum iteration entered by the user or the results achieved have converged (the new cluster center is the same as the old cluster center). If the position of the new centroids with the old centroids is not the same, then repeated (iteration).

Accurate method dispersed to monetary value can help formulating strategies to improve revenue while accurate dispersed to recency can help preventing churn out of customers [7]. In order to determine optimum number of k, we can combine some methods. Gustriansyah used eight indexes of validity to determine the optimal number of clusters namely Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index, which can improve the objectivity and similarity of data in product segmentation so that it can improve the accuracy of the stock management process [8]. Brahmana also used Davies Bouldin Index and Silhouette Index and implemented them to K-Means, K-Medoids, and DBSCAN algorithms [9].

RFM analysis is an approach to determine the nature of the customer in purchasing a product. This approach is very popular, especially in retail and industrial companies. The RFM model is a model for determining consumer clustering based on Recency, Frequency, and Monetary [10]. Recency is the period time (in units of days, months, years) from the last transaction made by consumers to date. Frequency is the total number of transactions or the average number of transactions in one period. Monetary is the average amount of costs incurred by consumers in a unit of time.

DESIGN SYSTEM

The analysis steps using the K-Means method begin with collecting distributor transaction history data. Next in the process of analyzing the K-Means method is to determine the value of K for the centroid, where the determination of K will be done randomly by the system. After the K value is obtained, a process is carried out to calculate the distance of each distributor transaction history data criteria to each centroid. The process of measuring the distance between the data and the center of the cluster used Euclidean distance [2].

The next step is to calculate the average of each centroid to determine the new centroid. The new cluster center is used to carry out the next iteration, if the results obtained have not converged. The iteration process will stop when it meets the maximum iterations entered by the user or the results achieved have converged (the new cluster center is the same as the old cluster center). If the position of the new centroids with the old centroids is not the same, then the

process is repeated until the maximum iteration. Figure 1 illustrates the flow of the distributor clustering analysis process using the K-Means method.

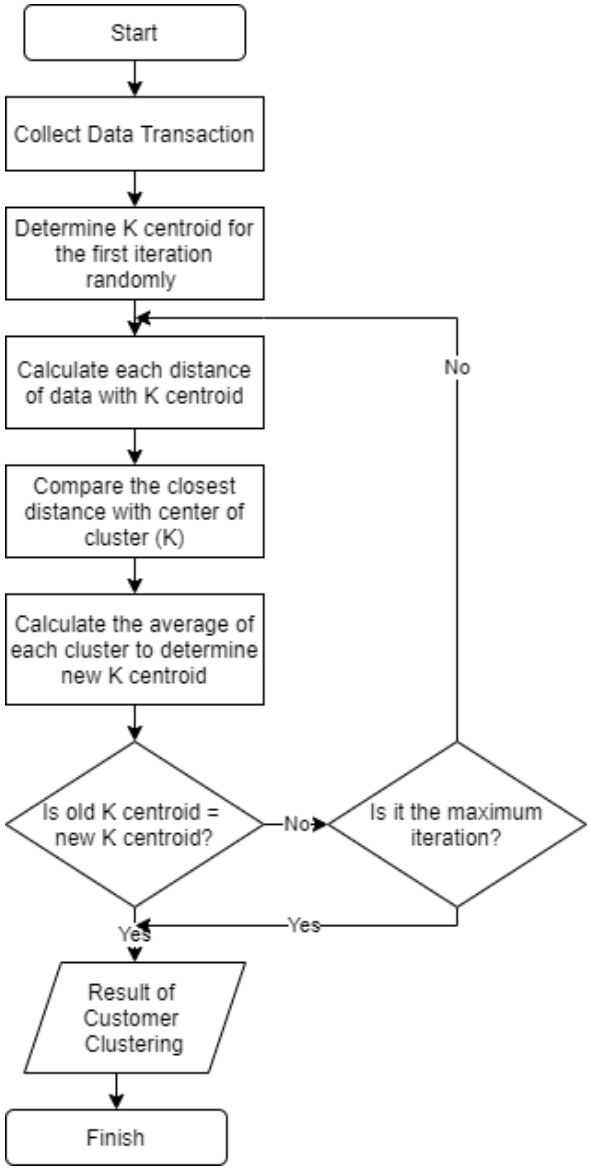

FIGURE 1. Algorithm of K-Means

IMPLEMENTATION

The calculation below only showing on the first 20 samples in 2019. The table of the purchase amount and the frequency can be seen in Table 1. The number will represent a distributor, the value of the purchase amount affected by recency, and frequency is the number of items purchased by the distributor.

After getting the value of the number of purchases and the frequency, the next step is to calculate the distance between each distributor and each centroid (the distance calculation for each class is not shown). The number of centroids to be used is six because the desired number of classes is 6, namely: Superstar, Golden, Typical, Occasional, Everyday, and Dormant. The closest distance to the centroid will determine the class of the distributor (class determined by looking at the nearest centroid is shown). The centroid point distance table can be seen in Table 2.

TABLE 1. The purchase amount and the frequency

Number	Purchase Amount (Affected by Recency)	Frequency
1	99,466	9
2	284,331,164	44
3	2,163,942,618	91
4	3,192,926,795	127
5	211,188,473	123
6	205,338,416	35
7	304,840,740	46
8	297,356,106	55
9	8,584,270,060	180
10	2,606,838,085	111
11	316,789,570	61
12	1,449,191,905	98
13	184,792,037	45
14	358,166,740	53
15	15,514,521	6
16	168,720,411	50
17	646,438	3
18	2,982,040	26
19	626,849	4
20	27,425	8

TABLE 2. The centroid point distance

Number	Purchase Amount (Affected by Recency)	Frequency	Classes
1	99,466	9	5
2	284,331,164	44	2
3	2,163,942,618	91	1
4	3,192,926,795	127	1
5	211,188,473	123	2
6	205,338,416	35	2
7	304,840,740	46	2
8	297,356,106	55	2
9	8,584,270,060	180	1
10	2,606,838,085	111	1
11	316,789,570	61	2
12	1,449,191,905	98	1
13	184,792,037	45	2
14	358,166,740	53	2
15	15,514,521	6	3
16	168,720,411	50	2
17	646,438	3	4
18	2,982,040	26	3
19	626,849	4	4
20	27,425	8	6

Furthermore, by looking at Table 3, the number of purchases and the frequency in each class will be averaged to produce a new centroid. The new centroid point table can be seen in Table 3.

TABLE 3. The centroid point distance

Classes	Purchase Amount (Affected by Recency)	Frequency
1	3,599,433,893	121
2	259,058,184	57
3	9,248,281	16
4	636,644	4
5	99,466	9
6	27,425	8

The previous step will be repeated (iteration) until the result of the previous iteration class and the current iteration do not change. The following is the final result of customer clustering after six iterations. The iteration stops at the 6th iteration because the 5th and 6th iterations produce the same class. The final result table of distributor clustering can be seen in Table 4.

TABLE 4. The final distributor clustering

Number	Purchase Amount (Affected by Recency)	Frequency	Classes
1	99,466	9	6
2	284,331,164	44	3
3	2,163,942,618	91	2
4	3,192,926,795	127	2
5	211,188,473	123	3
6	205,338,416	35	3
7	304,840,740	46	3
8	297,356,106	55	3
9	8,584,270,060	180	1
10	2,606,838,085	111	2
11	316,789,570	61	3
12	1,449,191,905	98	2
13	184,792,037	45	3
14	358,166,740	53	3
15	15,514,521	6	4
16	168,720,411	50	3
17	646,438	3	6
18	2,982,040	26	5
19	626,849	4	6
20	27,425	8	6

After the distributor is clustered, the class will be labelled with a name. The range of the number of purchases and the frequency of the class will be displayed. The table for giving levels to classes can be seen in Table 5.

TABLE 5. The label for each class

Classes	Purchase Amount (Affected by Recency)	Frequency	Label
1	8,584,270,060 - 8,584,270,060	180 - 180	Superstar
2	1,449,191,905 – 3,192,926,795	91 – 127	Golden
3	168,720,411 – 168,720,411	35 -123	Typical
4	15,514,521 – 15,514,521	6 – 6	Occasional
5	2,982,040 – 2,982,040	26 - 26	Everyday
6	27,425 - 646,438	3 - 9	Dormant

Admin from the system is able to view the K-Means result. He can confirm or adjust the recommendation of distributor clustering. While on the distributor dashboard page, they can see their membership status information (Fig.

2). On the quote page, distributors can view the quote data based on distributor clustering that has been entered by the sales division. The display of the Price Offer page can be seen in Figure 3.

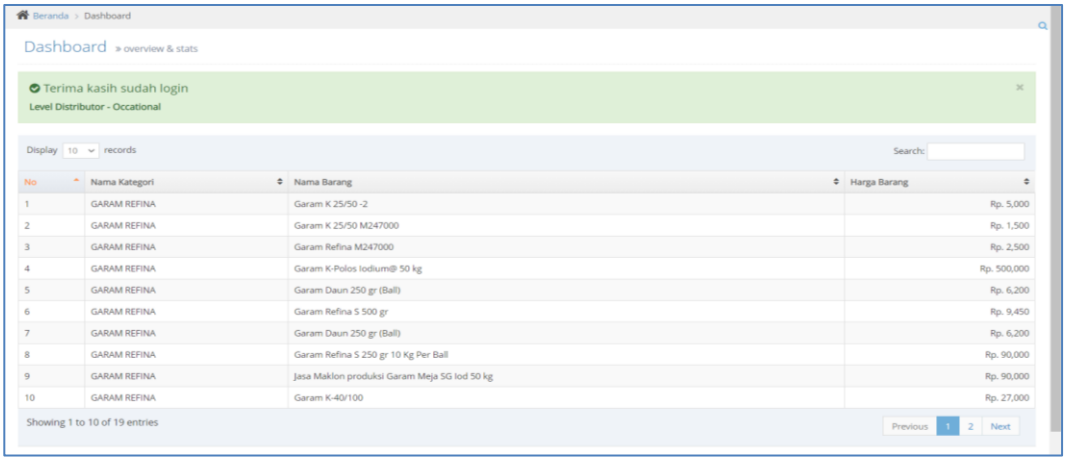

FIGURE 2. Dashboard of customer

FIGURE 3. Promotion for each clustered distributor

CONCLUSION

Based on the results of implementation, it can be concluded that the results of this study have succeeded in clustering distributor into six clusters to determine distributor level status with K-Means. For data in the year 2019, there are 509 active customers which are divided into six clusters as follows superstar cluster consists of six distributors, golden cluster consists of nine distributors, typical cluster consists of 19 distributors, occasional cluster consists of 12 distributors, everyday cluster consists of 22 distributors, and dormant cluster consists of 441 distributors. The last cluster is the default cluster for the average distributor. It also has succeeded in improving the CRM system to support the company's activities. The company could give special promotions based on the cluster.

REFERENCES

1. N. Bagul, P. Berad, P. Surana, and C. Khachane, "Retail customer churn analysis using RFM model and K-Means clustering", International Journal of Engineering Research & Technology, 10 (3), 2021, pp. 349-354.
2. P. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining* (Pearson, 2nd Edition, 2018).
3. K. G. Soni and A. Patel, "Comparative Analysis of K-means and K-medoids algorithm on IRIS data", International Journal of Computational Intelligence Research 13 (5), (2017), pp. 899-906.
4. P. Arora, Deepali and S. Varshney, "Analysis of K-Means and K-medoids algorithm for big data", Procedia Computer Science 78 (Elsevier, 2016), pp. 507-512.
5. B. Rizki, N. G. Ginasta, M. A. Tamrin, and A. Rahman, "Customer loyalty segmentation on point of sale system using Recency-Frequency-Monetary (RFM) and K-Means, Jurnal Online Informatika 5(2), (2020), pp. 130-136.
6. N. M. Taher, D. Elzanfaly, S. Salama, "Investigation in customer value segmentation quality under different preprocessing types of RFM attributes, International Journal of Recent Contributions from Engineering, Science & IT, 4 (4), (2016).

7. V. Dawane, P. Waghodekar, and J. Pagare, “RFM analysis using K-Means clustering to improve revenue and customer retention”, International Conference on Smart Data Intelligence (2021).
8. R. Gustriansyah, N. Suhandi, and F. Antony, “Clustering optimization in RFM analysis based on k-means”, Indonesian Journal of Electrical Engineering and Computer Science 18 (1), (2020), pp.470-477.
9. R. W. S. Brahmana, F. A. Mohammed, and K. Chairuang, “Customer segmentation based on RFM model using K-Means, K-Medoids, and DBSCAN methods”, Lontar Komputer 11 (1), 2020, pp. 32-43.
10. A. M. Hughes, *Strategic Database Marketing 4e: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program* (McGraw-Hill Education, 2012).