

# Data mining technique for grouping products using clustering based on association

Eka Praja Wiyata Mandala, Dewi Eka Putri

Department of Computer Science, University Putra Indonesia YPTK Padang, Padang, Indonesia

---

## Article Info

### Article history:

Received Nov 18, 2022

Revised Feb 16, 2023

Accepted Mar 12, 2023

---

### Keywords:

Association rules

Clustering

Grouping products

Hybrid data mining

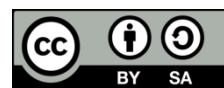
Minimarket

---

## ABSTRACT

There is high competition between these minimarkets so many products sold in each minimarket are not sold until they expire. The aim of this study is to help retail managers cluster products in minimarkets. The data obtained will be processed using the hybrid data mining approach by combining two methods in data mining. In the first section, association uses the FP-Growth algorithm, and in the second section, clustering uses the K-means algorithm. From the experimental results, it can be seen that the proposed approach can minimize the number of products to be grouped. After the association process is carried out, from 29 products in 12 transactions, 6 products can be obtained that has a frequency above the minimum support and minimum confidence. After the clustering process, 6 products are grouped into 2 clusters, so that 1 product is included in the most interested product cluster and 5 products are included in the interested product cluster. We minimize data processing so that retail managers can process data directly from sales transaction data on the cashier's computer and can quickly get the results of product grouping.

*This is an open access article under the [CC BY-SA](#) license.*



---

## Corresponding Author:

Eka Praja Wiyata Mandala

Department of Computer Science, Universitas Putra Indonesia YPTK Padang

Lubuk Begalung, Padang, 25221, Indonesia

Email: [ekaprajawm@upiyptk.ac.id](mailto:ekaprajawm@upiyptk.ac.id)

---

## 1. INTRODUCTION

At present, retail has made many changes from traditional stores with untidy product arrangements to shopping centers that are neat and well-managed. Therefore, these shops try to accommodate all items needed for daily use or items that are rarely needed under the same roof [1]. One example is a retail store that sells products, where the retail store is the party that has difficulty determining product sales. Included in this retail store category are minimarkets. The number of emergence of new minimarkets continues to increase, this indicates that minimarkets supply products that have not been fulfilled by supermarkets [2]. The products sold in minimarkets are many and varied, ranging from food, drinks, necessities, and so on. Sales transactions that occur in minimarkets are also very high, in a day can reach hundreds or even thousands of transactions. The products purchased by customers in each transaction are also not small so that the products sold from minimarkets in a day can also reach thousands of products.

Retail is an industry that is at the end of the production chain, where producers sell products in large quantities to retailers who will then resell them to customers who will use the product. The challenges faced by the retail industry occur continuously, so retail managers must work hard to survive. The development of digital sales with all the conveniences offered is starting to threaten the existence of retail stores. The main challenge faced by retail stores is the drastic decline in people's purchasing power, resulting in reduced sales from retail stores. The main cause is the emergence of online shopping trends so that customers can order products directly from home with only their cellphones and internet quota without having to bother going

shopping at retail stores. So, customer loyalty to the retail store will also decrease. As a result, the products sold by these retail stores do not sell and are not sold out until the expiration date of the product runs out.

The data mining approach can be used to solve the problem of product grouping at the retail store. Data mining is a method that is widely used in conducting research, where data mining can assist in solving problems that require analysis and can also solve business problems. Data mining capabilities in analyzing big data have been proven in various fields [3]. Data mining can be described as a process of finding hidden patterns and facts contained therein. Data mining can also be applied to business, marketing, detecting fraud and even data mining can be applied in the field of education [4]. Data mining has become an indispensable tool in the financial sector to transform data into meaningful information that can be used for making the best decisions to beat competitors who have the same line of business [5]. Data mining is one of the processes in KDD to extract patterns including association and clustering [6].

The data mining method that will be used in this paper is a hybrid data mining method, which combines two methods in data mining. Techniques for extracting data using hybrid data mining can provide more accurate predictions than independent data mining techniques. Hybrid data mining models will be very important to be used in data-scarce areas where technical skills and understanding of the processes that occur are still lacking [7]. Hybrid data mining is a combination of several methods, for example, a combination of several feature selection and classification learning algorithms that will be used for the decision-making process. Hybrid data mining will be made in several stages [8].

The methods that will be combined are the association method to obtain the products most frequently purchased by customers and the clustering method to group products at the retail store. Association rules are a very important technique for finding frequent patterns in data mining which are used for market basket analysis, computer networks, recommendation systems, and healthcare [9]. Association rules can find positive or negative relationships between different items and have received a lot of attention recently in many applications, such as web mining, recommender systems, and intrusion detection. Association rules are called valid rules if their support is greater than the user-defined minimum support threshold (ms) and the confidence is greater than the user-defined minimum confidence (mc) threshold [10]. Association rules can be used to explore possible relationships between precast production activities in a particular data set. Moreover, association rule learning is a rule-based machine learning method for finding relationships between individual items in a large data set [11]. Association rules have been studied to find the regularity between items in relational data. They have the traditional form  $X \Rightarrow Y$ , where  $X$  and  $Y$  are separate sets of items [12].

Clustering is a technique used in data mining in determining data groups or clusters so that they have the highest level of linkage between two data points if they are in the same cluster and have low linkages if they are in different clusters [13]. Clustering is a learning technique in data mining that is very important for many applications such as knowledge discovery, detecting process errors, recommendation systems, and anomaly detection [14]. Clustering is a very important tool in data mining, which helps in retrieving useful data from data that is available in very large quantities. Clustering is an approach that is considered effective where data can be grouped by looking at patterns or similarities between data in one data group [15]. Clustering is an unsupervised learning method used to group data sets so that the data items in a particular cluster are more similar than those in other clusters. The unsupervised learning model is used on data sets that are never labeled or classified [16]. Clustering can extract hidden features in datasets and can be classified into two methods, hierarchy and partitioning [17].

## 2. METHOD

In this study, we propose a hybrid data mining model to group goods in retail stores. In the first section, we searched for patterns of product sales frequency using the FP-Growth algorithm. In the second section, we group products based on the frequency pattern of product sales using the K-means algorithm, as shown in Figure 1.

Figure 1 illustrates the flowchart of the methodology proposed in this work. Our proposed approach is a two-section model. First, we generate product sales frequency patterns using the FP-Growth algorithm and then generate product grouping based on product sales frequency patterns using the K-means algorithm.

### 2.1. First section: FP-Growth algorithm

The FP-Growth algorithm is often used to find frequent itemsets, it cannot scale directly to current big data, especially for data sets that have large ranges. FP-Growth is based on the FP-Tree data structure, which stores item frequency information in a compact form. FP-Growth requires only two scans of the database and does not generate a pool of candidate items [18]. The FP-Growth algorithm is applied to detect patterns that often occur for each predetermined system problem. The FP-Growth algorithm is efficient and scalable to mine the full set of patterns that occur frequently with the growth of pattern fragments, using an extended prefix

tree structure to store compressed and important information about frequent patterns named frequent-pattern tree (FP-Tree) [19]. FP-Growth cancels the need for candidate generation. FP-Growth uses a tree-based data structure, named FP-Tree, to represent data sets in a concise manner that summarizes and mines frequent item sets from this structure [20].

The support value and confidence value are formulated as shown in (1) and (2) [21]:

$$Support(A) = \frac{\text{number of transactions containing } A}{\text{total transactions}} \tag{1}$$

$$Confidence(A, B) = \frac{\text{number of transactions containing } A \text{ and } B}{\text{number of transactions containing } A} \tag{2}$$

where *Support* is the percentage of those item combinations in the database and *Confidence* is the strength of the relationship between items in the formed associative rules.

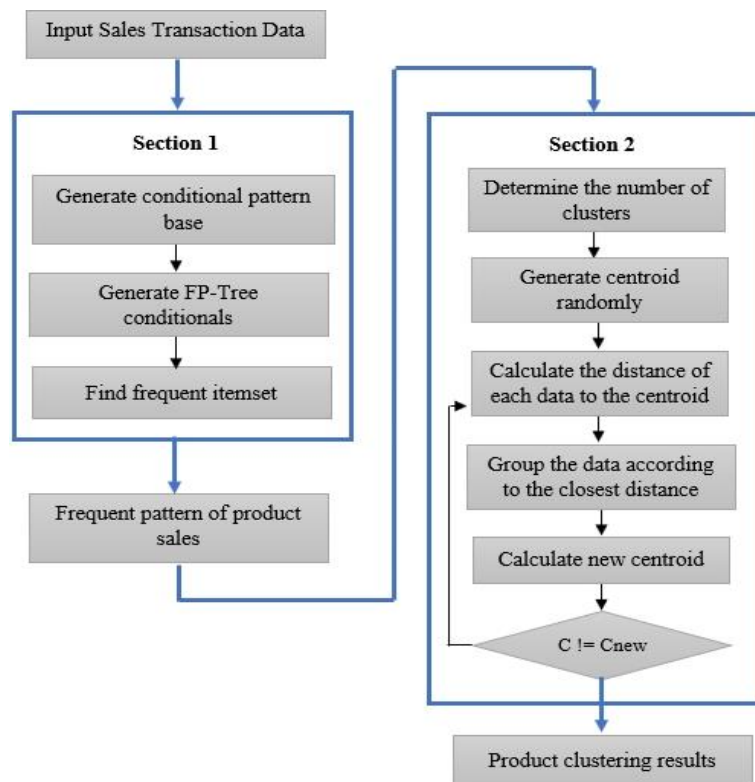


Figure 1. Research method

**2.2. Second section: K-means algorithm**

The K-means algorithm is an algorithm used in partitional clustering which was developed to group data sets with attributes consisting of numeric only. With different initialization, the K-means algorithm can generate different end clusters that are different as well [22]. The K-means algorithm is one of the most commonly used and well-studied clustering methods. K-means aims to divide all data points into cluster centers so that the number of intra-cluster variances can be minimized. K-means has four steps, namely; i) initialization, the number of cluster centers (K) is assigned and the initial cluster centers are randomly selected; ii) division, the data points are divided into the nearest cluster center according to the Euclidean pixel distance and the cluster center; iii) updating, the new cluster center is obtained by calculating the average reflection value of pixels; and iv) iterations, the process returns to step 2 until the cluster center does not change or the specified iteration number is reached [23].

In K-means, clusters are associated with a collection of points that appear around the cluster's centroid set. The optimal centroid is the one that minimizes the sum of the squared distances between each point and the specified centroid [24]. Clustering with the K-means algorithm is a grouping using the partition method, which is adapted to machine learning and pattern recognition problems. The K-means algorithm is very

sensitive to the initial centroid point which is calculated randomly for a given cluster [25]. To calculate the distance to the  $i$ -th data ( $x_{ij}$ ) at the center of the  $k$ -cluster ( $c_{kj}$ ) given the name  $d_{ik}$ , the Euclidean formula can be used [26].

$$d_{ik} = \sqrt{\sum_{j=1}^m (x_{ij} - c_{kj})^2} \quad (3)$$

Where  $d_{ik}$  is the distance between the data and the centroid,  $x_{ij}$  is the data coordinate and  $c_{kj}$  is the centroid coordinate.

Large datasets will be grouped with K-means to become several smaller groups. Appropriate product recommendations can help in marketing strategies, especially product promotion and production planning and helps in making decisions regarding product stock [27]. The clustering-based approach used not only consists of items that appear frequently but also considers their contribution to overall income by considering their prices [28].

### 3. RESULTS AND DISCUSSION

#### 3.1. Preprocessing

The data is taken from sales transactions at Sastra Mart. Sastra Mart is located in the Lubuk Begalung sub-district, Padang, West Sumatra. There are 12 transactions that have been filtered from the total transactions for one month. The transaction data used is only for transactions that contain many product items, a pattern of product sales frequency will be generated. There are 29 product items involved in 12 transactions which can be seen in Table 1.

Table 1. Product as data training

Product	Name	Initial
1	Beng Beng 20 gr	BE
2	Big Cola 1500 ml	BC
3	Chitato 68 gr	HT
4	Choco Mania 90 gr	CC
5	Chocolatos 200 ml	CB
6	Cimory 250 ml	CY
7	Delfi Cha Cha Peanut 10g	DE
8	Dua Kelinci 80 gr	DY
9	Fanta 1500 ml	FA
10	Floridina 350 ml	FF
11	Happy Tos 140 gr	HP
12	Kacang Atom Garuda 130gr	GR
13	Kapal Api 165 gr	KA
14	Koko Krunch 330 gr	CR
15	Kraft 165 gr	RA
16	Love 1000 ml	LX
17	Milo 30 gr	MA
18	Okky Koko Drink 195gr	OB
19	Oreo Vanila 130 gr	RE
20	Pocky 30 gr	PA
21	Relaxa 40 gr	RV
22	Sakatonik ABC 30 Tablet	SAC
23	Sari Roti Cokelat 72gr	SA
24	SGM 3 Vanilla 150gr	SG
25	Silver Queen 30 gr	SN
26	Sprite 1500 ml	ST
27	Teh Botol Sosro 450 ml	RO
28	Wafer Tango 130 gr	PTN
29	Walls Feast Vanilla 30 gr	WS

Table 1 defines the list of items involved from all transactions that will be obtained in this study. All of these items are spread across 12 product sales transactions. Data for all transactions to be processed can be seen in Table 2.

Table 2. Purchased product transactions

Transactions	Item Bought
1	CB, CY, HP, GR, FF, DE, DY, KA, CR, CC
2	LX, CB, DE, MA, FF, OB, GR, RE
3	GR, SN, DE, ST, CB, RO, MA, PTN
4	BE, BC, HT, CC, CB, CY, DE, DY, FA, FF
5	WS, RE, LX, CB, DE, MA
6	ST, DE, CB, RO, MA, PTN
7	CR, HP, GR, DE, DY, KA, CC
8	DE, CY, CR, HP, GR, PA, FF, RV
9	SG, DE, GR
10	SAC, DE, DY, KA, FA, LX
11	DE, DY, KA, LX, SA, FA
12	RA, CC, DE

3.2. Section 1: FP-Growth algorithm

This study uses the minimum support value, which is at least 1 item purchased 4 times or 33.33% of a total of 12 transactions. DE support value is obtained from the number of DE frequencies divided by total transactions.

$$Support_{(DE)} = \frac{\text{number of DE transactions}}{\text{transaction totals}} = \frac{12}{12} = 100\%$$

From the calculation results, obtained as many as 9 products that have a minimum support value greater than or equal to 4. Furthermore, each transaction will be sorted all the items from the largest support value to the smallest support value. Furthermore, a frequency tree can be created called the frequent-pattern tree (FP-Tree), as shown in Figure 2.

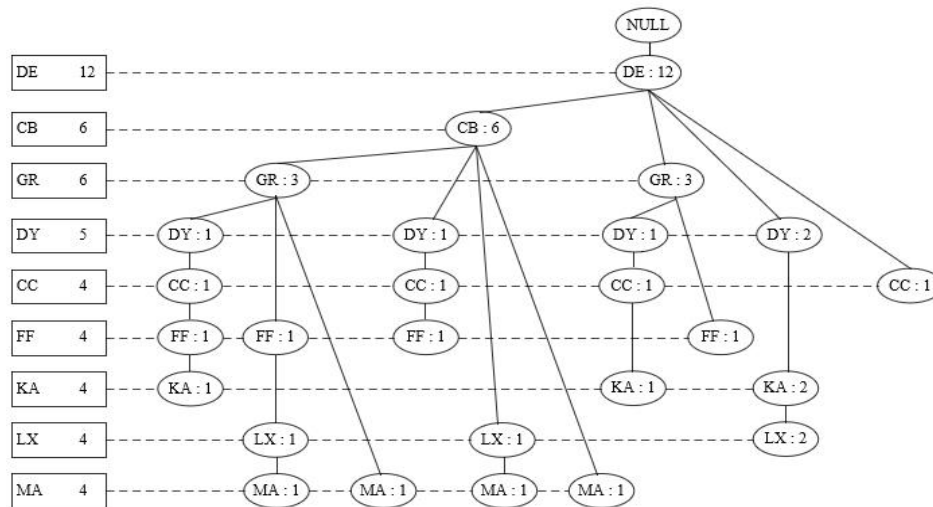


Figure 2. Frequent pattern tree

Figure 2 illustrates the data represented in the FP-Tree model. By using FP-Tree, the database scan is only done twice, no need to repeat it. FP-Tree is an excellent data structure for frequent pattern mining. This structure provides complete information to form a frequent pattern. There are 9 products that have a support value greater than or equal to the minimum support. There is 1 product that has a support value equal to 12 (100%), i.e. DE. There are 2 products that have a support value equal to 6 (50%), i.e. CB and GR. There is 1 product that has a support value equal to 5 (41.67%), i.e. DY. And there are 5 products that have a support value equal to 4 (33.33%) i.e. CC, FF, KA, LX, and MA.

The next step is to determine the conditional pattern base of all items that have a frequency equal to or greater than the minimum support. The conditional pattern base is a sub-database containing a path prefix and a pattern suffix. The basic generation of the conditional pattern is obtained through the FP-Tree. At this stage, the sum of the number of supports for each item in each conditional pattern base is carried out, then a

conditional FP-Tree is generated for each item that has the number of supports greater than or equal to the minimum number of supports. The next step is to find the frequent itemset, which is looking for a single path that is then combined with the items in the conditional FP-Tree. A frequent itemset is the number of transactions that contain a certain number of item sets. The point is the number of transactions that buy an item set. Frequent itemset can be seen in Table 3.

Table 3. Frequent itemset

Item	Frequent Itemset
MA	(DE → MA: 4), (CB → MA: 4), (DE, CB → MA: 4)
LX	(DE → LX: 4)
KA	(DE → KA: 4), (DY → KA: 4), (DE, DY → KA: 4)
FF	(DE → FF: 4)
CC	(DE → CC: 4)
DY	(DE → DY: 5)
GR	(DE → GR: 6)
CB	(DE → CB: 6)
DE	(DE → MA: 4), (CB → MA: 4), (DE, CB → MA: 4)

From the frequent itemset in Table 3, the confidence value can be calculated. Confidence is a comparison between the support value of the set of items contained in the rule and the support value of the set of items that precede it. Confidence can be obtained from the number of transactions containing DE and MA items divided by the total transaction items containing DE items.

$$Confident_{(DE,MA)} = \frac{\text{number of DE and MA transactions}}{\text{tnumber of DE transactions}} = \frac{4}{12} = 33,33\%$$

$$Confident_{(DE,GR)} = \frac{\text{number of DE and GR transactions}}{\text{tnumber of DE transactions}} = \frac{6}{12} = 50,00\%$$

The next step is to determine interesting rules (strong association rules). It must first determine the minimum confidence. If the minimum confidence value is set at 50%, then the interesting rules obtained are shown in Table 4. Table 4 provides interesting rules that meet the minimum support, which is 33.33%, and the minimum confidence, which is 50%, with 8 interesting rules. We tested the results of the calculations above to ensure the premise and conclusion values using RapidMiner, so that we obtained results as shown in Figure 3.

Table 4. Interesting rules

Interesting rules	Support	Confidence
If buy DE Then buy CB	50,00%	50,00%
If buy DE Then buy GR	50,00%	50,00%
If buy CB Then buy MA	33,33%	66,67%
If buy CB Then buy DE and MA	33,33%	66,67%
If buy DE and CB Then buy MA	33,33%	66,67%
If buy DY Then buy KA	33,33%	80,00%
If buy DY Then buy DE and KA	33,33%	80,00%
If buy DE and DY Then buy KA	33,33%	80,00%

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convict...
1	DE	CB	0.500	0.500	0.750	-1.500	0	1	1
2	DE	GR	0.500	0.500	0.750	-1.500	0	1	1
3	CB	MA	0.333	0.667	0.889	-0.667	0.167	2	2
4	CB	DE, MA	0.333	0.667	0.889	-0.667	0.167	2	2
5	DE, CB	MA	0.333	0.667	0.889	-0.667	0.167	2	2
6	DY	KA	0.333	0.800	0.941	-0.500	0.194	2.400	3.333
7	DY	DE, KA	0.333	0.800	0.941	-0.500	0.194	2.400	3.333
8	DE, DY	KA	0.333	0.800	0.941	-0.500	0.194	2.400	3.333

Figure 3. Test results with RapidMiner

Figure 3 shows the premises and conclusions generated using FP-Growth. This shows that the resulting 8 rules are accompanied by the values of support and confidence. Visualization of the rules obtained can be seen in Figure 4.

Figure 4 illustrates the visualization of the rules obtained from the FP-Growth process by displaying the support and confidence values of each rule. From the visualization results obtained eight rules which are divided into three groups of rules. The first group obtained a support value of 0.500 and a confidence value of 0.500. The second group obtained a support value of 0.333 and a confidence value of 0.667. And the last group obtained a support value of 0.333 and a confidence value of 0.800.

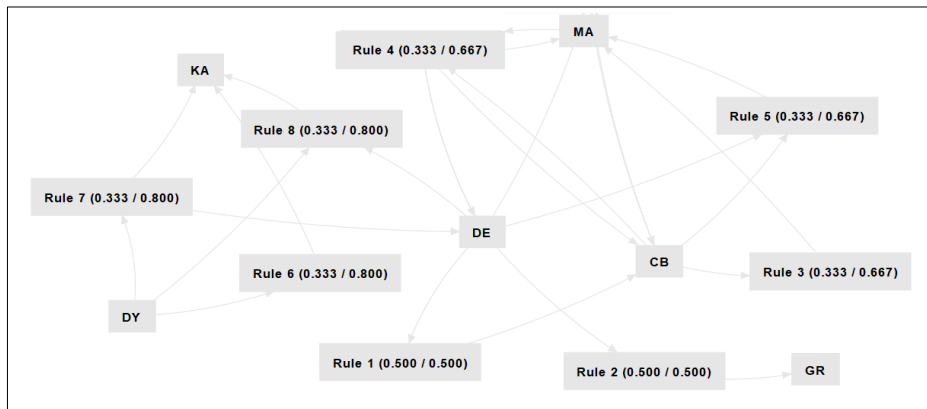


Figure 4. Visualization of the results with RapidMiner

**3.3. Section 2: K-means algorithm**

The following data used are taken from the results of calculations using the FP-Growth algorithm. There are 6 products involved in the formation of interesting rules in section 1. All products have a frequency greater than or equal to the minimum support that has been determined previously. It can be seen in Table 5.

Table 5. Product sales frequency

Item (m)	Frequency (x)
CB	6
MA	4
DE	12
DY	5
KA	4
GR	6

The products sold will be grouped into 2 clusters, the most interested product and the interested product. For the cluster with the most interested product (cluster 1), a centroid with a frequency of 9 will be generated. For the cluster with the interested product (cluster 2), a centroid with a frequency of 4 will be generated. So, the values of C1={9} and C2={4}. Calculate the distance between each item to each centroid using the correlation formula between two objects, Euclidean distance. The distance between each item (M) to C1 and C2 can be seen in Table 6.

Table 6. Item distance to C1 and C2

Cluster	M1	M2	M3	M4	M5	M6
C1	3.00	5.00	3.00	4.00	5.00	3.00
C2	2.00	0.00	8.00	1.00	0.00	2.00

The next step is to compare the distance of each item to C1 and C2. The distances that have been obtained will be grouped based on the shortest distance from each item to each centroid. For example, M1 is closer to C2 than to C1 so that M1 becomes a member of C2 as can be seen in Table 7.

Table 7. Member of each cluster

Cluster	Member
C1	M3
C2	M1, M2, M4, M5, M6

Then determine the new centroid value for each cluster by counting the number of each product selling frequency from each cluster. Because the C1new and C2new values are not the same as the C1 and C2 values, then the calculation of the distance from each item M to C1new and C2new is carried out so that the distance. The result is that there is no change in the members of each cluster. If no member changes occur, the process is terminated. So that the product sales groupings are obtained in Table 8.

Table 8. Grouping product

Member of most interested product	Member of interested product
M3 Delfi Cha Cha Peanut 10g	M1 Chocolatos 200 ml
	M2 Milo 30 gr
	M4 Dua Kelinci 80 gr
	M5 Kapal Api 165 gr
	M6 Kacang Atom Garuda 130gr

The grouping results in Table 8 can be displayed visually using a scatter plot. The use of scatter plots aims to display and clarify the location of the data. The data displayed is seen to be divided into two clusters where the grouping results for each cluster can be seen in Figure 5.

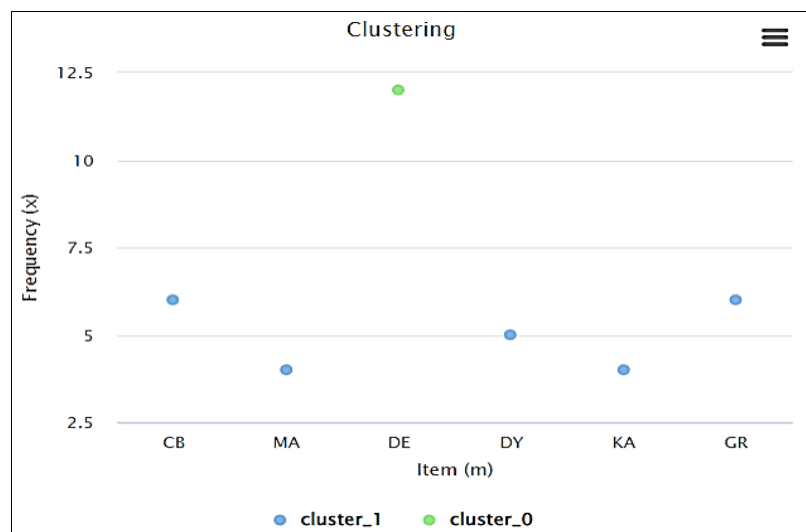


Figure 5. Scatter plot of the clustering results with RapidMiner

We measure cluster validity on a clustering method using the davies-bouldin index (DBI). DBI is used to minimize the inter-cluster distance and at the same time try to minimize the distance between points in a cluster. We obtained measurement results with a DBI of 0.114. With this DBI value, it can be concluded that the similarity of data in one cluster is 88.6%.

#### 4. CONCLUSION

In this study, the problem of grouping products sold has been discussed. A hybrid data mining approach is proposed in product grouping. The process begins with finding a pattern of product sales frequency, then grouping products based on sales frequency. From the experimental results, it can be seen that the proposed approach can minimize the number of products to be grouped. After the association process is carried out, from 29 products in 12 transactions, 6 products can be obtained that has a frequency above the minimum



support and minimum confidence. After the clustering process, 6 products are grouped into 2 clusters, so that 1 product is included in the most interested product cluster and 5 products are included in the interested product cluster. The results of testing using DBI, we found the similarity of data in one cluster was 88.6%. In this study, we minimize data processing from a large number of sales transaction data to obtain product clusters that are sold, so that retail managers can process data directly from sales transaction data on the cashier's computer and can quickly get product sales grouping results. The hybrid data mining method introduced in this study can be used with large datasets, but requires a long processing time. The impact of this research for business actors is to make it easier for business actors to determine the next product procurement and assist in stock management. For customers, the impact of this research is that customers will find it easier to find the products they need so customer loyalty increases. Some of the limitations of the study include requiring tools in data processing so that the process can be directly processed from transaction data into product groupings. In the future, we will try to classify product sales from the results of the previous product clustering.

## ACKNOWLEDGEMENTS

We would like to thank the Chairperson of the Yayasan Perguruan Tinggi Komputer (YPTK) Padang, Mrs. Dr. Hj. Zerni Melmusi, SE, MM, Ak, CA for providing us with the opportunity and funding this University Development Research with contract number: 6/UPI-YPTK/LPPM/KP/PGB/I/2021. We also thank the Rector of Universitas Putra Indonesia YPTK Padang, Prof. Dr. H. Sarjon Defit, S.Kom., M.Sc. who have provided opportunities and learning as well as guidance to researchers.




## REFERENCES

- [1] N. Verma, D. Malhotra, and J. Singh, "Big data analytics for retail industry using MapReduce-Apriori framework," *Journal of Management Analytics*, vol. 7, no. 3, pp. 424-442, 2020, doi: 10.1080/23270012.2020.1728403.
- [2] A. P. Graciola, D. D. Toni, G. S. Milan, and L. Eberle, "Mediated-moderated effects: high and low store image, brand awareness, perceived value from mini and supermarkets retail stores," *Journal of Retailing and Consumer Services*, vol. 55, no. April, p. 102117, 2020, doi: 10.1016/j.jretconser.2020.102117.
- [3] S. Jain and V. Kumar, "Garment categorization using data mining techniques," *Symmetry (Basel)*, vol. 12, no. 984, pp. 1-20, 2020, doi: 10.3390/sym12060984.
- [4] S. J. Ghorpade, S. S. Patil, and R. S. Chaudhari, "Educational data mining: tools and techniques study," *International Journal of Research and Analytical Reviews (IJRAR)*, vol. 7, no. 4, pp. 70-82, 2020.
- [5] M. Ibrahim and D. O. Tayo, "Data mining: theory, concept and techniques," *International Journal of Science and Innovative Research*, vol. 5, no. 2, pp. 17-26, 2020.
- [6] P. Bertalya, L. Setyowati, F. I. Irawan, and S. R. Irianti, "Formulation of city health development index using data mining," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 1, pp. 362-369, 2021, doi: 10.11591/ijeecs.v23.i1.pp362-369.
- [7] K. Khosravi, J. R. Cooper, P. Daggupati, B. T. Pham, and D. T. Bui, "Bedload transport rate prediction: Application of novel hybrid data mining techniques," *Journal of Hydrology*, vol. 585, p. 124774, 2020, doi: 10.1016/j.jhydrol.2020.124774.
- [8] J. Nalić, G. Martinović, and D. Žagar, "New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers," *Advanced Engineering Informatics*, vol. 45, p. 101130, 2020, doi: 10.1016/j.aei.2020.101130.
- [9] A. Telikani, A. H. Gandomi, and A. Shahbahrami, "A survey of evolutionary computation for association rule mining," *Information Sciences*, vol. 524, pp. 318-352, 2020, doi: 10.1016/j.ins.2020.02.073.
- [10] X. Dong, F. Hao, L. Zhao, and T. Xu, "An efficient method for pruning redundant negative and positive association rules," *Neurocomputing*, vol. 393, pp. 245-258, 2020, doi: 10.1016/j.neucom.2018.09.108.
- [11] J. H. Chen, S. C. Hsu, C. L. Chen, H. W. Tai, and T. H. Wu, "Exploring the association rules of work activities for producing precast components," *Automation in Construction*, vol. 111, p. 103059, 2020, doi: 10.1016/j.autcon.2019.103059.
- [12] X. Wang, Y. Xu, and H. Zhan, "Extending association rules with graph patterns," *Expert System Application*, vol. 141, pp. 1-16, 2020, doi: 10.1016/j.eswa.2019.112897.
- [13] P. Borthakur and B. Goswami, "Short term load forecasting: a hybrid approach using data mining methods," *In 2020 International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, 2020, pp. 1-6, doi: 10.1109/ICEFEET49149.2020.9187009.
- [14] J. Maia *et al.*, "Evolving clustering algorithm based on mixture of typicalities for stream data mining," *Future Generation Computer Systems*, vol. 106, pp. 672-684, 2020, doi: 10.1016/j.future.2020.01.017.
- [15] S. Singh and S. Srivastava, "Review of clustering techniques in control system," *Procedia Computer Science*, vol. 173, pp. 272-280, 2020, doi: 10.1016/j.procs.2020.06.032.
- [16] M. Cui, "Introduction to the K-means clustering algorithm based on the elbow method," *Accounting, Audit, Finance*, vol. 1, pp. 5-8, 2020, doi: 10.23977/accf.2020.010102.
- [17] Y. Sato, K. Izui, T. Yamada, and S. Nishiwaki, "Data mining based on clustering and association rule analysis for knowledge discovery in multiobjective topology optimization," *Expert System Application*, vol. 119, pp. 247-261, 2019, doi: 10.1016/j.eswa.2018.10.047.
- [18] S. Bagui, K. Devulapalli, and J. Coffey, "A heuristic approach for load balancing the FP-Growth algorithm on MapReduce," *Array*, vol. 7, no. July, p. 100035, 2020, doi: 10.1016/j.array.2020.100035.
- [19] J. Wang and Z. Cheng, "FP-Growth based regular behaviors auditing in electric management information system," *Procedia Computer Science*, vol. 139, pp. 275-279, 2018, doi: 10.1016/j.procs.2018.10.268.
- [20] L. Shabtay, P. Fournier-Viger, R. Yaari, and I. Dattner, "A guided FP-Growth algorithm for mining multitude-targeted item-sets and class association rules in imbalanced data," *Information Sciences*, vol. 553, pp. 353-375, 2021, doi: 10.1016/j.ins.2020.10.020.
- [21] J. R. Chang, Y. S. Chen, C. K. Lin, and M. F. Cheng, "Advanced data mining of SSD quality based on FP-growth data analysis," *Application Science*, vol. 11, no. 4, pp. 1-15, 2021, doi: 10.3390/app11041715.




- [22] A. Ahmad and S. S. Khan, "initKmix-A novel initial partition generation algorithm for clustering mixed data using K-means-based clustering," *Expert System Application*, 2019, p. 114149, 2020, doi: 10.1016/j.eswa.2020.114149.
- [23] Z. Ren, L. Sun, and Q. Zhai, "Improved k-means and spectral matching for hyperspectral mineral mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 91, p. 102154, 2020, doi: 10.1016/j.jag.2020.102154.
- [24] D. P. Hofmeyr, "Degrees of freedom and model selection for K-means clustering," *Computational Statistics and Data*, vol. 149, p. 106974, 2020, doi: 10.1016/j.csda.2020.106974.
- [25] S. Manochandar, M. Punniyamoorthy, and R. K. Jeyachitra, "Development of new seed with modified validity measures for K-means clustering," *Computers and Industrial Engineering*, vol. 141, p. 106290, 2020, doi: 10.1016/j.cie.2020.106290.
- [26] H. Bian, Y. Zhong, J. Sun, and F. Shi, "Study on power consumption load forecast based on K-means clustering and FCM-BP model," *Energy Reports*, vol. 6, pp. 693–700, 2020, doi: 10.1016/j.egy.2020.11.148.
- [27] M. Imron, U. Hasanah, and B. Humaidi, "Analysis of data mining using K-means clustering algorithm for product grouping," *International Journal of Informatics and Information Systems (IJIS)*, vol. 3, no. 1, pp. 12–22, 2020, doi: 10.47738/ijis.v3i1.3.
- [28] S. Kanhere, A. Sahni, P. Stynes, and P. Pathak, "Clustering based approach to enhance association rule mining," *In 2021 28th Conference of Open Innovations Association (FRUCT)*, vol. 2021-Janua, 2021, doi: 10.23919/FRUCT50888.2021.9347577.

## BIOGRAPHIES OF AUTHORS



**Eka Praja Wiyata Mandala**    obtained a Bachelor's degree in Computer Science majoring in Informatics Engineering at Esa Unggul University, Jakarta, Indonesia. Master's degree in Computer Science majoring in Informatics Engineering was obtained from the Universitas Putra Indonesia YPTK, Padang, Indonesia. Currently pursuing a Doctoral Program in Information Technology at the Universitas Putra Indonesia YPTK, Padang, Indonesia by taking database science concentration. His research interests include data mining, fuzzy logic, expert system, and artificial neural network. He can be contacted at email: [ekaprajawm@upiptk.ac.id](mailto:ekaprajawm@upiptk.ac.id).



**Dewi Eka Putri**    obtained a Bachelor's degree in Computer Science majoring in Information System at Esa Unggul University, Jakarta, Indonesia. Master's degree in Computer Science majoring in Informatics Engineering was obtained from the Universitas Putra Indonesia YPTK, Padang, Indonesia. Her research interests include data mining, fuzzy logic, and artificial neural network. She can be contacted at email: [dewieka@upiptk.ac.id](mailto:dewieka@upiptk.ac.id).