IJASER

International Journal of Applied Science and Engineering Review

ISSN: 2582-6271

Vol. 3, Issue.6, Nov-Dec 2022, page no. 43-65

To cite this article: Singgih Saptadi, Ary Arvianto, Wiwik Budiawan and Dhimas Wachid Nur Saputra (2022). PRODUCT CLUSTERING IN THE MSME BUSINESS OF GROCERY STORE, International Journal of Applied Science and Engineering Review (IJASER) 3 (6): 43-65

PRODUCT CLUSTERING IN THE MSME BUSINESS OF GROCERY STORE

Singgih Saptadi, Ary Arvianto, Wiwik Budiawan and Dhimas Wachid Nur Saputra

Diponegoro University, Industrial Engineering Department,
Semarang, Indonesia

DOI: http://dx.doi.org/10.52267/IJASER.2022.3604

ABSTRACT

The business world is an exciting world to follow due to its dynamic and competitiveness. Sri Wahyuni Grocery is one of the MSMEs involved in buying and selling daily-basis needs. The business can manage an average of 115 transactions in a day, including various transactions for necessities. Most products purchased at this grocery shop are daily basic needs, such as sugar, tea, instant food, snacks, and fuel oil. It is known that from observations, some products are less desirable by buyers so they are not selling well and some products are selling well, so it is necessary to do a product grouping process to find out how to group these products. An analysis of existing data is needed to obtain information with the K-Means Clustering algorithm. This research aims to determine the pattern of transaction data owned by Traditional Grocery Store MSMEs and form a Clustering pattern of products offered by Traditional Grocery Store MSMEs. Through the data exploration process, this research will carry out a pattern search from transaction data and clustering patterns owned by Traditional Grocery Store MSMEs. Based on the findings from the research conducted, business actors will be able to develop various strategies to improve services to sales by relying on the use of various data mining algorithms. The research was conducted on Traditional Grocery Store MSMEs with transaction data for two months, and the research carried out data exploration to determine clustering patterns using the CRISP-DM method.

KEYWORDS: MSMEs, Grocery Store, Data Mining, K-Means Clustering, RapidMiner.

1. INTRODUCTION

The business world is exciting to follow due to its dynamic and competitiveness among existing business actors [13]. Grocery stores' MSMEs directly compete with minimarkets. These two types of retail businesses provide a variety of daily needs. According to data from Euromonitor International in Indonesia, the number of minimarkets has recorded a reasonably high increase and is at 39% from 2015 to 2020. This high number was also accompanied by the number of outlets that continued to experience an increasing trend [1]. The existence of modern minimarkets and modern retail indeed threatens the impact on traditional grocery store entrepreneurs. The impact is a decrease in turnover to a decrease in the number of buyers [2].

Sri Wahyuni Grocery Store is one of the MSMEs engaged in the business sector of buying and selling basic needs of the community. Located in Jatipuro Village, Karanganyar, this shop has a building area of 5 meters × 7.5 meters. The store can serve transactions in a day with an average of 115 transactions containing various types of transactions for necessities.

According to observations and interviews with parties related to consumer behavior who shop at Sri Wahyuni Stores, consumers or buyers tend to purchase goods intended for daily needs, such as various basic needs, sugar, tea, various types of instant food, and snacks [3]. In addition, buyers at this grocery store also tend to come from multiple areas in Jatipuro District, ranging from children to elderly adults. Most of these buyers or consumers have an income of IDR 1,800,000 – IDR 3,000,000 per month. In addition to basic needs, another product that is also one of the best-selling products is fuel oil or BBM. This shop can sell up to 140 liters of fuel per day, from Pertalite to Solar.

Clustering algorithms can provide knowledge for preparing various strategies that can help traditional grocery stores thrive. It is known that from observations, some products are less in demand by buyers so that they are not sold well in sales, then some products sell well, so it is necessary to carry out a product grouping process to find out how the product grouping is. Vladimír Holy et al. (2017) conducted a study to classify products into several clusters according to their appearance in the shopping cart of buyers [4]. Furthermore, there is a study entitled Retail business analytics: Customer visit segmentation using market basket data in 2018 with the purpose of the study is to find out more deeply and understand the shopping behavior and intentions of customers on each transaction and allow retailers to adjust the needs of customers appropriately. Therefore, the study was conducted with clustering techniques to identify customer visit segmentation. Briefly, it can be explained that this study was carried out by segmenting customer visits and then related to the customers' purchase intentions behind the holidays. For example, visitors buy biscuits, chocolates, drinks, ice cream, snacks, and chips in the data segment. Then it can be concluded that this customer intends to purchase snacks and beverages [5].

Customers' purchase decisions are known to be influenced by various factors. The factors include the quality of the items to those offered at a discount. These two actions exemplify a marketing technique that business actors can use [6]. On the other case, customers tend to feel comfortable shopping at minimarkets or modern retail because of the services provided, the facilities offered, to the layout of product arrangements that are attractive to them [2]. These various factors can be studied using data mining, which will help develop various aspects of the strategy [7]. However, determining the strategy cannot be done immediately. It is necessary to have a process of understanding and extracting information to prevent the possibility of the adverse impact of the strategy to be drawn up. So based on this explanation, this research formulates the research question as follows.

RQ1: What is the pattern of transaction data owned by Traditional Grocery Store MSMEs?
RQ2: How do Traditional Grocery Store MSMEs own the product clustering pattern?
This study adopted the CRISP-DM model to find transaction data patterns and create a Clustering model for obtaining basic knowledge, which can later be developed into formulating various strategies to develop a business.

2. LITERATURE REVIEW

2.1. Data Mining Process

The data mining process is divided into several main processes: understanding the problem, preparing sample data, making models, applying models to datasets to see how the model will work in the real world, and deploying and maintaining existing models. More details can be seen in the following figure 1 [8].

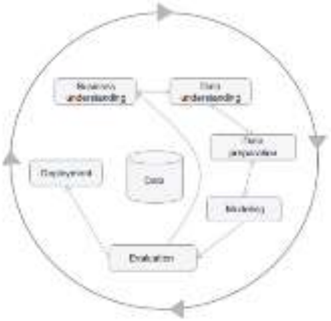


Figure 1: Data Mining Process

CRISP-DM, or Cross Industry Standard Process for Data mining, is one of the most popular processes in data science. The framework was developed by an association of companies involved in data mining. CRISP-DM is usually used to provide solutions to problems that are done with the help of data science [8].

2.2. K-Means Clustering Algorithm

A clustering algorithm is one of the processes by which a set of data is broken down into several subsets. Each subset that exists is called a Cluster with a state where each data in the same Cluster has similar traits or characteristics while other data that does not have different traits or characters will be in another subset or Cluster. Cluster analysis is commonly used in business problems, recognition of depiction patterns, web search, biology, and security [12]. We will discuss the usefulness of the clustering algorithm in the business problem sector. In this sector, clustering can be used to group customer data. Each group has characteristics, and the data has the same characteristics. Clustering becomes very useful to help develop business strategies to improve customer relationships. In addition to connecting with customers, clustering can be used in grouping projects into various categories based on the similarity of audits and project diagnosis so that the project can be carried out effectively [8].

3. APPROACH TO DATA EXPLORATION

This type of research using exploration data has been done before. Data exploration has previously been the subject of research that included both application and literature research. Christoph Schröer et al. (2021) conducted a study with a systematic literature review on the application of the CRISP-DM model. In his research, Christoph reviewed the CRISP-DM model used in the latest study to provide findings related to research focus, best practices to innovative methods [9]. Then there is research from Veronika Plotnikova et al. (2022) conducted research by trying to apply CRISP-DM to the financial services industry as one of the methods for carrying out the data mining process. The research results identify the consistency in the CRISP-DM lifecycle, impacts, and solutions to overcome gaps if used the CRISP-DM model in the financial service industry [10].

This research took the object of research in the form of one of the Micro, Small, and Medium Enterprises or MSMEs of the Grocery Store "Sri Wahyuni" located in Jatipuro, Karanganyar. Data collection was carried out for two months in January and February. Then the research method is selected using the CRISP-DM method. Here are the stages of research conducted.

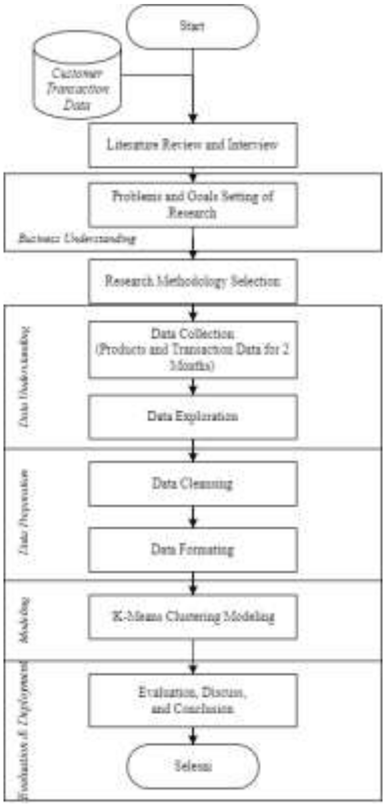

Start
Customer Transaction Data
Literature Review and Interview
Problems and Goals Setting of Research
Business Understanding
Research Methodology Selection
Data Understanding
Data Collection (Products and Transaction Data for 2 Months)
Data Exploration
Data Preparation
Data Cleansing
Data Formating
Modeling
K-Means Clustering Modeling
Evaluation & Deployment
Evaluation, Discuss, and Conclusion
Selesai

Figure 2: Research Methodology

3.1. Data Understanding

3.1.1. Data Collection

Data collection is carried out directly on the object of study. The data taken is data on products sold and purchase transactions from customers. Data collection in the form of customer purchase transactions is taken for two months, namely January and February. Here is an example of a product data:

Table 1: Products Data

Product Type	Product Brand
Cigarette	Dji Sam Soe Refill
	Dji Sam Soe
	Sampoerna Kretek
	Magnum Filter
	...
	Magnum Blue
...	...
Body Care	Dettol
	Shinzui
	Citra Sabun Mandi
	Imperial Leather
	Nuvo Family
	Lifebuoy
	...
	Flacenta

Here is an example of customer transaction data.

Table 2: Customer Transaction Data

ID	Quantity of Goods	Type of Product	Price	Total Price	Brand
1	1	Pertalite	Rp8.500,00	Rp8.500,00	
2	1	Pertamax Turbo	Rp13.000,00	Rp15.000,00	
	1	Snack	Rp2.000,00		Simba Cereal
3	1	Gas	Rp16.000,00	Rp64.000,00	
	2	Gula	Rp28.000,00		
	3	Tea	Rp18.000,00		Dandang
	1	Milk	Rp2.000,00		Milo
4	1	Pertalite	Rp8.500,00	Rp10.500,00	
	1	Snack	Rp2.000,00		Tic Tic
5	2	Cigarette	Rp18.000,00	Rp22.000,00	Andalan
	1	Lighter	Rp2.000,00		
	4	Coffee	Rp2.000,00		ABC Plus
6	1	Gas	Rp16.000,00	Rp16.000,00	
…	…	…	…	…	…
6962	1	Cigarette	Rp13.000,00		Menara

3.1.2. Data Exploration

Data exploration is carried out on each attribute before the data is processed using data mining techniques. The following table 1 is the results of data exploration related to revenue and the number of customers from grocery stores with descriptive statistics in January.

Table 3: January Descriptive Statistic Exploration

Daily Income in January		*Customer in January*	
Mean	2819983.871	Mean	122.45161
Standard Error	83382.48599	Standard Error	2.9277459
Median	2806000	Median	121
Mode	#N/A	Mode	120
Standard Deviation	464254.034	Standard Deviation	16.300999
Sample Variance	2.15532E+11	Sample Variance	265.72258
Kurtosis	0.936062872	Kurtosis	2.4553677
Skewness	-0.21619275	Skewness	-0.3989151
Range	2290000	Range	90
Minimum	1594000	Minimum	73
Maximum	3884000	Maximum	163
Sum	87419500	Sum	3796
Count	31	Mean	122.45161

Table 3 shows that in the daily income attribute for January, the average income obtained by the grocery store is IDR 2.819.983,871 with a total of stores open for 31 days in the month. The revenue range in January is IDR 2.290.000 - with the maximum revenue obtained by the store in a day IDR 3.884.500 while the lowest income in January is IDR 1.594.000. Next, for the January transaction attribute, it is known that in-store shoppers averaged 123 transactions in a day for 31 days after opening the store. The range of customers making transactions in stores is as many as 90, with most transactions in a day reaching 163 customers, and the least in a day as many as 73 transactions. In January, a total of 3796 transactions were carried out at this grocery store and provided a total income to the store of Rp. 41.740.500, -. Then it was discovered that the data has an extensive range, so it varies. Then it is known from the standard deviation value owned by the data, where this standard deviation value is smaller than the mean value. Then, this existing data does not show the presence of outliers or extreme data. Based on the results of skewness and kurtosis, the daily income attribute is normally distributed in the range of -2 to 2. The existing data tend to lean to the right because the skewness value is positive, more than 0, and the existing data is homogeneous by being indicated by the positive kurtosis value. According to the results of skewness and kurtosis of the transaction attributes, the daily income attribute is known to be an abnormal distribution because it is not in the range of -2 to 2, with the tendency of the existing data to lean toward the left. After all, the skewness value is negative less than 0, and the existing data is homogeneous by being indicated by the positive kurtosis value. Moving on to February, table 2 is the results of the exploration of data related to revenue and the number of customers from the grocery store with descriptive statistics in February.

Table 4: February Descriptive Statistic Exploration

Daily Income in February		*Customer in February*	
Mean	2874196.429	Mean	113.14286
Standard Error	103474.1211	Standard Error	1.8557178
Median	2845500	Median	115
Mode	#N/A	Mode	116
Standard Deviation	547533.5833	Standard Deviation	9.8195357
Sample Variance	2.99793E+11	Sample Variance	96.42328
Kurtosis	-0.611417217	Kurtosis	1.1030068
Skewness	0.185282322	Skewness	-0.297317
Range	1949000	Range	46
Minimum	1922500	Minimum	90
Maximum	3871500	Maximum	136
Sum	80477500	Sum	3168
Count	28	Count	28

In February, it can be known that the first attribute is the average daily income earned by the store is Rp2.874.196,429, - with the total duration of the store opening during the month being 28 days. The range of income earned by the store is Rp1.949.000, - with the maximum opinion obtained by the store in a day being Rp3.871.500,- while for the lowest income in February it is Rp1.922.500,-. Next, for the second attribute, transactions in February, the average transaction made at this grocery store is 114 in 28 days of operation. The range of customers making transactions in stores is 46, with the most transactions in a day reaching 136 and the least in a day as many as 90 transactions. In February, there were 3168 transactions carried out at this grocery store which provided a total income to the store of Rp33.936.000, -. Then it can be known that the data has an extensive data range, so the data there varies. Then judging from the standard deviation value owned by the data, where this standard deviation value is smaller than the mean value of the data, this existing data does not show the presence of outliers or extreme data. According to the results of skewness and kurtosis, it is known that the two attributes are normally distributed in the range of -2 to 2, with the tendency of the data on the daily income attribute lean to the right because the skewness value is positive, more than 0 and the data tends to be the width of the bottom, indicated by the negative kurtosis value. Meanwhile, the existing transaction attribute is skewed to the left because the skewness value is negative less than 0. The existing data is homogeneous, as indicated by the positive kurtosis value.

The following is a comparative graph of earnings from January and February.

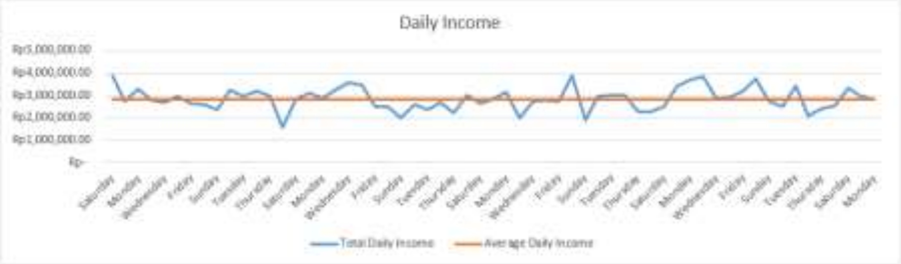


Figure 3: Daily Income Comparison

From the graph in figure 3, the data pattern shown is a stationary type of pattern. The data pattern looks up and down but is still around the average of the known amount of income. The highest income was on Saturday of the first week, Rp3.884.000, - and the lowest was on Friday of the second week, Rp1.594.000, -. If the diagram draws a straight line with the average income obtained of Rp2.845.711, - then it can be said that this income data has a stationary data pattern type, with the distribution of data around the average income during the two-month data collection process.

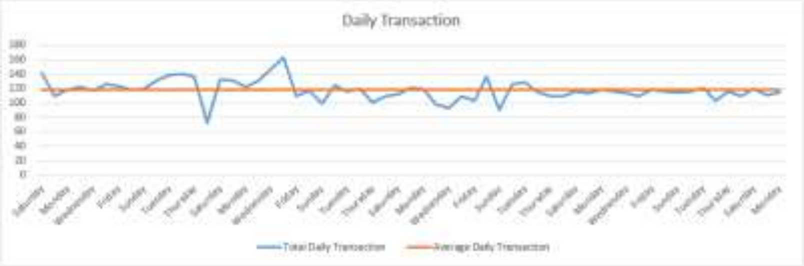


Figure 4: Transaction Comparison

From the chart in figure 4, it's known that the transaction data forms a stationary data pattern type, this can be known from the data where the distribution of data seems to have increased and decreased, but this increase and decrease is still around the average total number of transactions per day. The most transactions in a day occurred on Thursday in the third week, to be precise, in January, 163 transactions, while the least number of transactions occurred on Friday in the second week to be precise in January also with a total of 73 transactions.

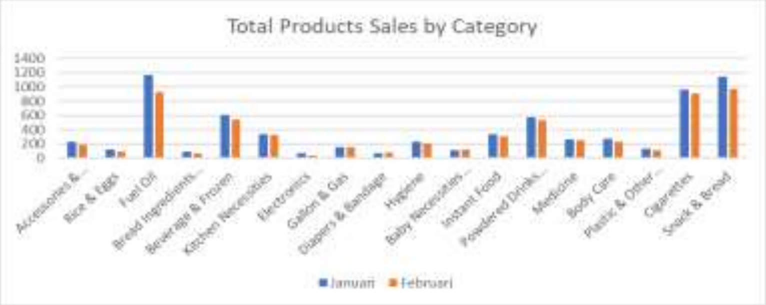


Figure 5: Comparison of Sales from Each Category

It can be seen in figure 5, where the comparison of sales of goods in each category is quite extreme. In January, the category with the highest sales was Fuel Oil Category, and this category was sold in as many as 1169 categories during 31 working days. Furthermore, it also happened in February, when Fuel Oil was the category with the highest product sales compared to products from other categories within 28 working days of 928 categories of goods. Meanwhile, the category with the minor product sales in January the Diapers & Bandage category, where this category only sold 67 categories of goods, and for February, the minor sales were the Electronics category, which only sold as many as 34 categories of goods. For the difference between the two months, it is known that the most significant difference between the two months is in the Fuel Oil Category, with the resulting difference of 241 categories of goods with an advantage in January. Almost overall, January leads the sales results. However, there are some Categories where February leads, such as in the Diapers & Bandage category with a margin of 9 and Baby Necessary Category with a resulting margin of 12. Next will be seen the sales results within two months, judging from per pcs of products from each category. Here is a histogram chart.

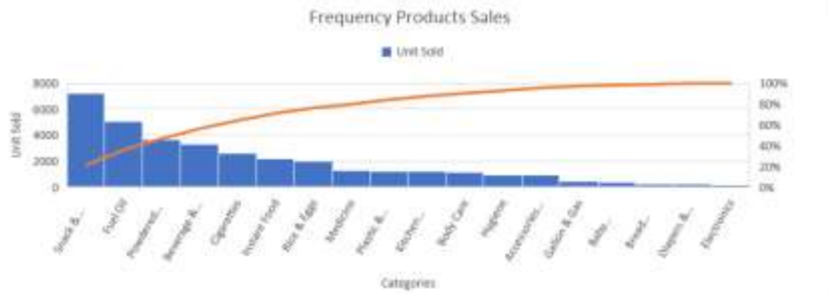


Figure 6: Frequency Products Sales

Based on the histogram chart shown in figure 6, two product categories look more prominent than the other categories. These two categories have a higher product sales frequency than other product categories. These two categories are the Fuel Oil category and the Cigarette category. Each of these categories sold

2097 transactions for Fuel Oil and 1875 cigarettes for cigarettes. The graph also shows such an extreme difference if a product from one category is very marketable, the sales become very high, while other products lag in this sale. The category with the minor frequency of sales is the Rice category which only sells as many as 31 product transactions for two months. The difference formed is very high, reaching 2066 transactions. Next, the Other Needs category can only sell as many as 32 product transactions, meaning that the difference is only about 1 product transaction between the Rice Category and the Powder Drink Category. That further shows the extreme difference in the frequency of sales transactions of each category, where the two highest categories have a difference of 222 product transactions. The bottom two categories have a difference of 1 product transaction, but a difference or a significant difference occurs between the top category and the lowest category, which is 2066 product transactions.

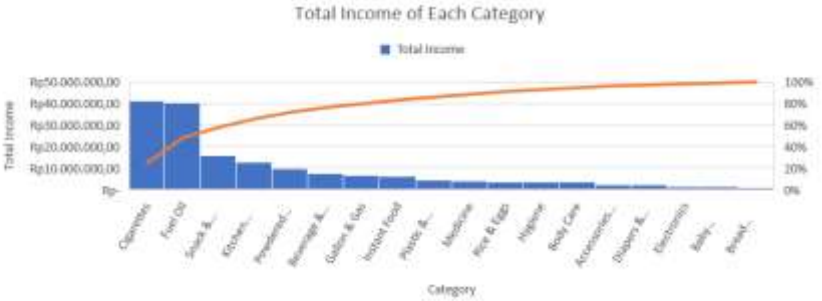


Figure 7: Amount of Revenue from Each Category

Based on the histogram chart shown in figure 7, knowing that two categories stand out in the amount of income earned by the store, the difference looks extreme. These two categories are Fuel Oil and Cigarettes. The Fuel Oil category managed to sell 5086 liters of products with revenues of Rp40.522.000, - and the second category of Cigarettes managed to sell products as much as 2631 pcs with revenues of Rp41.222.500,-. The income of the Cigarette category is higher than the Fuel Oil category with the difference in income of these two categories of Rp700.500,-. That can happen because the products offered by the Cigarette category are much more diverse with various prices that are also diverse compared to the fuel oil category products. For example, the price of one cigarette product at Rp30.000 for each pack, but with Rp30.000, customers can buy only 3 liters of Pertamax fuel oil.

Furthermore, from the results of the previous exploration regarding the number of units sold, the Snack & Candy category, which in the previous exploration was ranked second highest in unit sales in this search, actually ranked third with total revenue of Rp10.343.500,-. This number is far from the income earned by the Fuel Oil category. So, this condition shows that both the price and quantity produced in sales are significant to pay attention to be considered by business actors. Categories that sell little do not necessarily generate less income. The important thing is how the strategy must solve the problem so that both factors that are part of the income factor of a business are carried out. The Fuel Oil, Cigarette, and Snack & Candy categories remain the top three of the two factors. However, the Other Needs Category, which occupies the bottom two in product unit sales, is no longer included in the bottom three based on store income. Based on table 9, we need a validation that there are categories where the categories of products offered can be divided into two Clusters using the Clustering algorithm.

3.2. Data Preparation

3.2.1. Data Cleansing

Data cleansing is performed on both existing datasets. In this process, no noisy and incomplete data were found, so the data was ready to go to the formatting stage. Data cleansing can be seen by checking the data on the RapidMiner application, as shown figure 8 below.

Figure 8: Data Cleansing

3.2.2. Data Formatting

Data formatting aims to change existing data format by creating a grouping hierarchy that will later create categories of each existing product. Here is a hierarchy of product groupings for Grocery Stores, as shown in figure 9 below

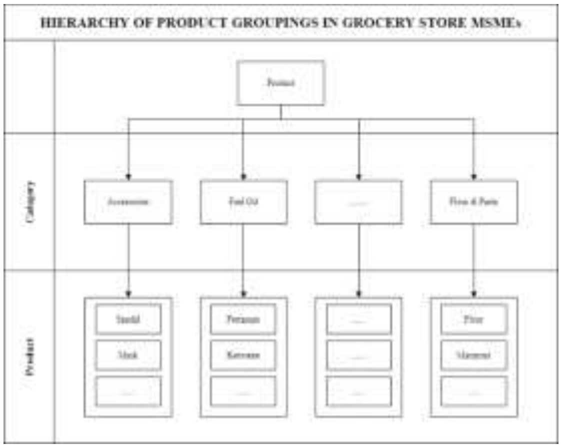


Figure 9: Product Hierarchy

After grouping, the data will be converted into a binary format according to the category or group. The following is an example of changing or formatting data for use in product category clustering modeling.

Table 5: Data Clustering Format

No	Category	Unit Sold	Transaction	Total Income
1	Accessories & Stationery	890	433	2296500
2	Rice & Eggs	1951	210	3659000
3	Fuel Oil	5086	2097	40522000
4	Bread Ingredients & Flour	228	171	1136500
5	Beverage & Frozen	3286	1196	7598000
6	Kitchen Necessities	1193	727	12763000
7	Electronics	173	103	1456500
8	Gallon & Gas	430	313	6772000
9	Diapers & Bandage	224	143	2061000
10	Hygiene	911	443	3574500
11	Baby Necessities & Toys	344	236	1428000
12	Instant Food	2170	639	6042500
13	Powdered Drinks & Milk	3633	1177	9504000
14	Medicine	1302	513	3973000
15	Body Care	1132	579	3519500
16	Plastic & Other Necessities	1209	235	4352000
17	Cigarettes	2631	1875	41222500

No	Category	Unit Sold	Transaction	Total Income
18	Snack & Bread	7205	2517	16031500

3.3. K-Means Clustering Modeling

Clustering models are created to group data or cluster on data according to the characteristics' similarity. Grouping will make data with the same properties into one group or cluster itself. Modeling was performed with the K-Means Clustering Algorithm. Modeling begins with determining the attributes to be used. Several attributes were used, such as the attributes of the number of product sales, the number of transactions containing the product, and the total income from each category. After that, the value of k, or the number of Clusters used in the modeling process, is determined. Determination of the value of k uses the formation of a hierarchy of clusters with the following results as shown figure 10.

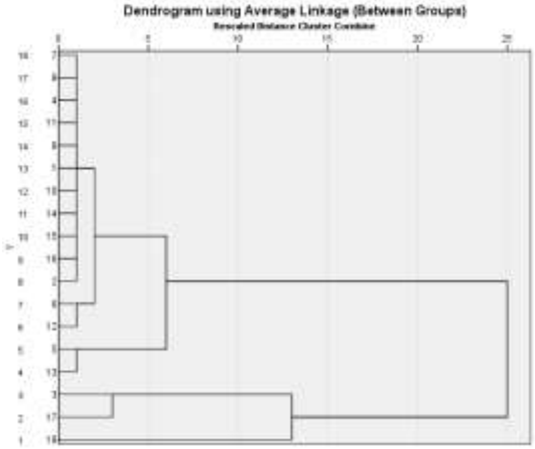


Figure 10: Dendrogram Hierarchy Clustering

The optimal k value may be utilized at k = 2 or k = 3 according to the dendrogram in figure 10. The value of k = 2 was used for this research. The following step is to normalize the data with the same scale as the other data.

3.3.1. Manual Modeling

Data normalization is the first process in the clustering modeling process. After calculating the normalization of data, the results are known in the following table.

Table 6: Data Normalization

No	Category	Unit Sold	Transaction	Total Income
1	Accessories & Stationery	-0.5429	-0.4565	-0.5951
2	Rice & Eggs	0.0338	-0.7716	-0.4798
3	Fuel Oil	1.7380	1.8955	2.6397
4	Bread Ingredients & Flour	-0.9028	-0.8268	-0.6932
5	Beverage & Frozen	0.7595	0.6220	-0.1464
6	Kitchen Necessities	-0.3782	-0.0409	0.2906
7	Electronics	-0.9327	-0.9229	-0.6662
8	Gallon & Gas	-0.7930	-0.6261	-0.2163
9	Diapers & Bandage	-0.9049	-0.8663	-0.6150
10	Hygiene	-0.5315	-0.4423	-0.4869
11	Baby Necessities & Toys	-0.8397	-0.7349	-0.6686
12	Instant Food	0.1529	-0.1653	-0.2781
13	Powdered Drinks & Milk	0.9481	0.5951	0.0149
14	Medicine	-0.3190	-0.3434	-0.4532
15	Body Care	-0.4114	-0.2501	-0.4916
16	Plastic & Other Necessities	-0.3695	-0.7363	-0.4211
17	Cigarettes	0.4035	1.5817	2.6990
18	Snack & Bread	2.8898	2.4891	0.5672

The following process is the determination of the initial centroid or central centroid. The determination of the initial centroid is carried out randomly. In this study, the data used as the initial centroid is the category of instant food and eggs. Furthermore, the first iteration calculation will be carried out from this central centroid. Here is an example of a calculation for iteration in the K-Means Clustering process.

Iteration C1 Accessories & Stationery :

$$D_e = \sqrt{((x_1 - s_1)^2 + (x_2 - s_2)^2 + (x_3 - s_3)^2}$$

(1)

$$= \sqrt{((0{,}4035 - (-0{,}5429))^2 + (1{,}5817 - (-0{,}4565))^2 + (2{,}699 - (-0{,}5951))^2} = 3{,}9875$$

Iteration C2 Accessories & Stationery :

$$D_e = \sqrt{((x_1 - s_1)^2 + (x_2 - s_2)^2 + (x_3 - s_3)^2}$$
$$= \sqrt{(-0{,}5429 - (-0{,}5429))^2 + (-0{,}4565 - (-0{,}4565))^2 + (-0{,}5951 - (-0{,}5951))^2} = 0$$

Then the first iteration process is carried out, with results as shown table 7.

Table 7: Iteration 1

No	Category	C1	C2	Cluster	Closest Distance	WCV
1	Accessories & Stationery	3.9875	0.0000	Cluster 2	0.0000	0.0000
2	Rice & Eggs	3.9723	0.6673	Cluster 2	0.6673	0.3336
3	Fuel Oil	1.3722	4.6041	Cluster 1	1.3722	0.6861
4	Bread Ingredients & Flour	4.3605	0.5256	Cluster 2	0.5256	0.2628
5	Beverage & Frozen	3.0240	1.7495	Cluster 2	1.7495	0.8747
6	Kitchen Necessities	3.0073	0.9921	Cluster 2	0.9921	0.4961
7	Electronics	4.4025	0.6120	Cluster 2	0.6120	0.3060
8	Gallon & Gas	3.8477	0.4845	Cluster 2	0.4845	0.2422
9	Diapers & Bandage	4.3229	0.5472	Cluster 2	0.5472	0.2736
10	Hygiene	3.8885	0.1097	Cluster 2	0.1097	0.0548
11	Baby Necessities & Toys	4.2723	0.4135	Cluster 2	0.4135	0.2068
12	Instant Food	3.4609	0.8182	Cluster 2	0.8182	0.4091
13	Powdered Drinks & Milk	2.9111	1.9238	Cluster 2	1.9238	0.9619
14	Medicine	3.7635	0.2882	Cluster 2	0.2882	0.1441
15	Body Care	3.7682	0.2657	Cluster 2	0.2657	0.1329
16	Plastic & Other Necessities	3.9630	0.3723	Cluster 2	0.3723	0.1862
17	Cigarettes	0.0000	3.9875	Cluster 1	0.0000	0.0000

18	Snack & Bread	3.3985	4.6702	Cluster 1	3.3985	1.6992

The iteration must continue until the cluster state stabilizes, where the entire category has no switching clusters. The stable state in the study was achieved in the second iteration. So that the clustering results are obtained as follows.

Table 8: Clustering Result

No	Category	Cluster
3	Fuel Oil	Cluster 1
17	Cigarettes	Cluster 1
18	Snack & Bread	Cluster 1
1	Accessories & Stationery	Cluster 2
2	Rice & Eggs	Cluster 2
4	Bread Ingredients & Flour	Cluster 2
5	Beverage & Frozen	Cluster 2
6	Kitchen Necessities	Cluster 2
7	Electronics	Cluster 2
8	Gallon & Gas	Cluster 2
9	Diapers & Bandage	Cluster 2
10	Hygiene	Cluster 2
11	Baby Necessities & Toys	Cluster 2
12	Instant Food	Cluster 2
13	Powdered Drinks & Milk	Cluster 2

No	Category	Cluster
14	Medicine	Cluster 2
15	Body Care	Cluster 2
16	Plastic & Other Necessities	Cluster 2

3.3.2. RapidMiner Modeling

Several operators are needed in carrying out RapidMiner modeling, including data operators, normalization operators, Clustering K-Means operators, multiply operators and visual model cluster operators. With an arrangement like the following figure 11.

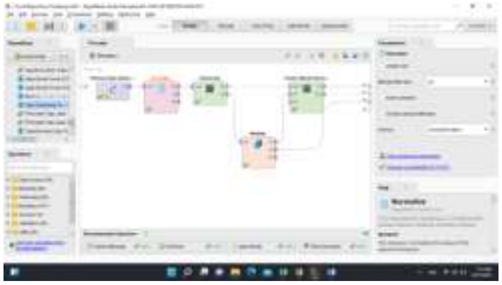


Figure 11: K-Means Clustering RapidMiner

Based on the modeling, the same results as manual modeling were obtained, as in the following figure 12 and figure 13.

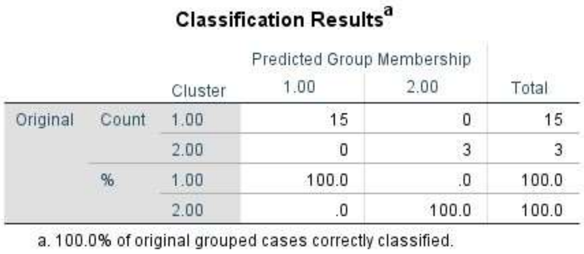
Cluster Model

Figure 12: Visualization Result of K-Means Clustering RapidMiner

A total of 18 category data were used. From the 18 categories of data used, two Clusters were formed with the number of each different Cluster. Cluster 1 consisted of 3 data members, while cluster 2 consisted of

15 data members. Each Cluster has its characteristics. Obtained from the output of the RapidMiner application, Cluster 1 has a total income with a more excellent average value of Rp97.776.000 compared to Cluster 2, which has an average value of the amount of income of Rp 70.136.000. Then the second attribute seen from the number of transactions in Cluster 1 is more excellent. The average value of the number of transactions from this Cluster is 2163 greater than Cluster 2, which is smaller and is at an average value of 475. Last seen in the third attribute, Cluster 1 is more excellent, with the average value of the number of units sold 4974 greater than Cluster 2, which is at an average value of 3633. According to this characteristic, those Cluster with three members are Clusters with products that are in demand and widely purchased by customers as Cluster 1. In contrast, Cluster with fifteen members is a Cluster with products that are less in demand and are still rarely purchased by customers as Cluster 2.

Validation related to clustering results with two clusters was performed using the SPSS Discriminant Classification model. The result is shown in figure 13.



Classification Results[a]

			Predicted Group Membership		
		Cluster	1.00	2.00	Total
Original	Count	1.00	15	0	15
		2.00	0	3	3
	%	1.00	100.0	.0	100.0
		2.00	.0	100.0	100.0

a. 100.0% of original grouped cases correctly classified.

Figure 13: Clustering Validation

Figure 13 shows that the Cluster model produces a value of 100%, which means that the modeling has produced an appropriate classification.

4. RESULT AND DISCUSS

Transaction data owned by MSME players at Sri Wahyuni Grocery Store shows the pattern of transactions among customers. Data taken during the 2-month research process showed that sales of products showed a stationary pattern type with sales that always went up and down around the average in terms of daily income attributes and the number of transactions made with varying data distribution. In January, transaction attribute data were abnormally distributed with the tendency for existing data to be skewed to the left and existing data to be homogeneous. Meanwhile, in February, the transaction attribute data were normally distributed. The data leaned to the left because the value of skewness is homogeneous.

Transactions at traditional grocery stores tend to be dominated by single-item transactions. In this situation, one sales transaction only contains one item purchased by the customer. Based on the modeling, in traditional grocery stores, the grouping of products can be carried out on the scale of category divisions. This situation results from the traditional grocery store's purpose, which is to serve the daily requirements of the community, as opposed to modern retail businesses, which do serve this purpose by fulfilling monthly purchasing. Additionally, the findings of the clustering modeling contain three categories in Cluster 1 and fifteen categories in Cluster 2.

Regarding the amount of data, modern retail has more data that can be mined using data mining algorithms than traditional grocery stores. However, that does not mean that traditional grocery stores cannot use the same algorithm to develop and continue to compete in the business world. So, from the discovery of initial knowledge using data exploration and clustering algorithms, business actors can develop it into various strategies to cover shortcomings or lags in several aspects, such as price and profitability, promotional tools, and improve store layouts and arrangement of products [11].

5. CONCLUSION

5.1. Conclusion

The result of this study is that the sales pattern of MSMEs at Sri Wahyuni traditional grocery store has a stationary pattern, where the income from this transaction has increased and decreased which is still between the average income. Then, we performed modeling using the Clustering algorithm to find clustering patterns from these MSMEs. Modeling was made because of indications of significant differences in sales and income, so modeling with this algorithm is needed to help find out the Cluster pattern of these Traditional Grocery Store MSMEs. The mining clustering process starts with a hierarchy of product category divisions, and the following process is the mining clustering process. The clustering process divides the existing categories into two clusters with the results of three product categories included in Cluster 1 as products that are in great demand by buyers consist of the categories Fuel Oil, Snack & Bread, and Cigarette. Cluster 2 defined as the products that are less in demand by buyers, consists of the categories Accessories & Stationery, Rice & Eggs, Bread Ingredients & Flour, Beverage & Frozen, Kitchen Necessities, Electronics, Gallon & Gas, Diapers & Bandage, Hygiene, Baby Necessities & Toys, Instant Food, Powdered Drinks & Milk, Medicine, Body Care, and Plastic & Other Necessities.

5.2. Future Research

Researchers can conduct research using the product division level with more comprehensive historical data of transactions to see whether association rules may be formed at any given time. Then researchers can use or compare more models to get better results. Based on the knowledge obtained, further research can be carried out by providing or compiling various strategies to help business actors in several traditional grocery stores.

REFERENCES

[1] R. Pahlevi, "Jumlah Gerai Minimarket Meningkat 39% pada 2020," 6 July 2021. [Online]. Available: https://databoks.katadata.co.id/.

[2] R. Masruroh, "The impact of modern retail Minimarket towards the continuity of traditional retail Businesses," *IOP Conference Series: Materials Science and Engineering (Vol. 180, No. 1, p. 012005)*, pp. 1-6, 2017.

[3] S. Sunanto, "Modern retail impact on store preference and traditional Retailers in West Java," *Asian Journal of Business Research*, p. 2(2), 2012.

[4] V. Holy, O. Sokol and M. Cerny, "Clustering retail products based on customer behaviour," *Applied Soft Computing 60*, pp. 752-762, 2017.

[5] A. Grivia, C. Bardaki, K. Pramatari and D. Papakiriakopoulos, "Retail business analytics: Customer visit segmentation using market basket data," *Expert Systems with Applications*, pp. 1-16, 2018.

[6] "Amin, S.; Mahasan, S. S.," *Difference between consumer preferences to choose between the traditional retailing and modern retailing.*, pp. 63-70, 2019.

[7] J. Han, J. Pei and M. Kamber, Data Mining: Concepts and Techniques, Cambridge: Morgan Kaufmann, 2022.

[8] B. Deshpade and V. Kotu, Data Science Concepts and Practice 2nd Edition, Cambridge: Elsevier, 2019.

[9] C. Schröerab, F. Kruse and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science 181*, pp. 526-534, 2021.

[10] V. Plotnikova, M. Dumas and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data & Knowledge Engineering 139*, p. 102013, 2022.

[11] H. Chaniago, "Investigation of Factors Influencing Traditional Retail Success in Small Cities in Indonesia," *Journal of Applied Economic Sciences*, pp. 65-75, 2020.

[12] J.-L. Seng and T. Chen, "An analytic approach to select data mining for business decision," *Expert Systems with Applications*, pp. 8042-8057, 2010.

[13] V. S. Moertini, "Data Mining Sebagai Solusi Bisnis," *Integral Vol. 7*, pp. 44-56, April 2002.

Author Profile

Singgih Saptadi holds bachelor's, master's and doctoral degrees in industrial engineering, from the Bandung Institute of Technology. Since 2001, he has been a lecturer at the Department of Industrial Engineering, Diponegoro University and joined the Decision Support System Laboratory. Much of his research is in the implementation of decision-making and information technology in various industries.