IJISCS (International Journal of Information System and Computer Science)
IJISCS, Volume 9, Number 1, April 2025
E ISSN: 2598-246X; P ISSN: 2598-0793, pp. 1-9
Accredited SINTA 4 Number 79/E/KPT/2023
https://jurnal.ftikomibn.ac.id/index.php/ijiscs

APPLICATION OF DATA MINING FOR CLUSTERING THE USE OF TRACTOR VEHICLE SPAREPART UNITS USING THE K-MEANS ALGORITHM

Anggi Dwi Satriawan[1], Rustam[2], Nurmayanti[3], Sigit Mintoro[4], Supriyanto[5], Dwi Marisa Efendi[6]
Faculty Computer Science, Institute of Business Technology and Language, Lampung, Indonesia
Jl. Raya Candi Mas, Kotabumi, Lampung Utara, Lampung
Email : anggisatriawan02@gmail.com, rustam@dcc.ac.id, nurmayanti@dcc.ac.id, sigit@dcc.ac.id, supriyanto@dcc.ac.id, dwimarisa@dcc.ac.id

*Corresponding author
Email:
anggisatriawan02@gmail.com,
rustam@dcc.ac.id

Article history:
Received February 3, 2025
Revised February 26, 2025
Accepted March 11, 2025

Keywords:
Clustering;
data mining;
k-means.

Abstract
Inefficient management of tractor spare parts inventory can lead to high storage costs and operational downtime. This issue demands effective calculations to group spare parts based on usage patterns and procurement time. This research aims to apply data mining techniques to cluster tractor spare parts usage using the k-means algorithm to optimize inventory management. The methodology used involves data on spare parts usage over two years, which is then processed using the k-means algorithm to form several clusters based on usage frequency and lead time. This algorithm groups spare parts into clusters that minimize within-cluster variance and maximize between-cluster variance. The formed clusters are interpreted to determine the level of importance of the spare parts and their implications for inventory management strategies. The expected result is the identification of five main clusters grouping spare parts based on usage patterns with very high, medium, and low usage, as well as different lead time variations. These findings are expected to provide important insights for developing more efficient stock management strategies, reducing inventory costs, and increasing the availability of spare parts that match the operational needs of tractors, thus supporting overall efficiency in spare parts usage.

1.0 INTRODUCTION

In the manufacturing industry, machine maintenance with routine servicing and replacement of spare parts is a crucial aspect in maintaining the smoothness and productivity of a tool and extending the life of the machine. Efficient spare parts management plays an important role in optimizing machine performance, reducing the risk of equipment failure, and

providing operational efficiency to avoid customer dissatisfaction. The optimal use of spare parts can help organizations avoid stock shortages that may disrupt operational processes or services, which would undoubtedly disappoint customers as it could lead to delays in production processes and cause losses to the company. Conversely, overstocking can lead to unnecessary storage costs. Therefore, a deep understanding of spare part usage patterns can provide valuable insights for management in efficiently managing inventory.

In recent years, data mining techniques have become essential tools for analyzing and understanding hidden patterns in large datasets. One popular data mining technique is clustering, which aims to group data into several clusters so that data within a cluster have maximum similarity while data between clusters have minimum similarity based on certain attributes. In the context of spare parts inventory management, clustering techniques can be used to identify similar usage patterns of spare parts, allowing companies to categorize these spare parts into different groups and manage their inventory more efficiently.

One clustering algorithm that can be used is the k-means algorithm. This algorithm has several advantages: it is easy to implement and run, relatively fast, easy to adapt, and widely practiced in data mining tasks[1]. Previous research using the k-means algorithm includes grouping the best-selling products, clustering smokers over the age of 15, and grouping regions based on population density in an area, where the results can be considered as a basis for preventing social problems due to overpopulation[2],[3],[4]. The algorithm works by grouping data into several k clusters, where each data point is attributed to the cluster whose center is closest to it. K-means has proven effective in various applications, including clustering spare part usage patterns.

In this context, this research aims to apply data mining techniques, specifically the k-means algorithm, to analyze the usage patterns of spare parts needed for tractor units. By implementing this approach, it is expected that management will gain a better understanding of the usage patterns of spare parts, particularly for tractor units, and manage inventory more efficiently, thereby enhancing operational efficiency and customer satisfaction. This research is also expected to contribute to the advancement of knowledge in the fields of inventory management and data mining techniques, as well as provide practical recommendations for organizations in optimizing the use of spare parts.

2.0 THEORETICAL

2.1. Data Mining

Data mining is a mining process that uses statistical, mathematical, and artificial intelligence techniques to identify important and useful information from a large database[5]. Data mining is also defined as an activity that describes an iterative analysis process on large databases, aiming to extract accurate and potentially useful information and knowledge for knowledge workers involved in decision-making and problem-solving[6]. Data mining is often considered similar to KDD (Knowledge Discovery in Databases), but it is actually just one component of KDD. Compared to KDD, data mining is more popular among business practitioners[7].

2.2. Rapidminer

Rapidminer is an open-source software for knowledge discovery and data mining. It includes approximately 400 data mining procedures (operators), encompassing operators for input, output, data preprocessing, and visualization[8]. RapidMiner is a standalone application written in Java, allowing it to run on all operating systems[9].

2.3. Elbow Method

The Elbow method is used to select the optimal number of clusters. The Elbow algorithm is used to determine the number of groups to be formed[10]. The Elbow method is implemented by determining the optimal data and observing the graph of the assigned k-values[11].

2.4. K-Means Algorithm

Clustering is a process of grouping data into several clusters or groups such that the data within a cluster have maximum similarity and minimum dissimilarity[3]. K-means is a data mining algorithm that falls under unsupervised learning, meaning that the algorithm groups the input data without prior knowledge of the target classes or labels. This is different from supervised learning data mining, where the classes or labels are predetermined and the algorithm is

trained accordingly. K-means clustering divides the data set into k clusters, where each data point is assigned to one cluster with a centroid representing the average of the data within that cluster[12]

K-means partition data points into k clusters based on the distance metric used for grouping. The value of 'k' must be specified by the user. The distance is calculated between data points and the cluster centroids. Data points closest to the cluster centroid are assigned to that cluster. After each iteration, it recalculates the centroids of the clusters, and this process continues until the predetermined number of iterations is reached or until the cluster centroids no longer change after an iteration[13].

The steps in implementing the k-means algorithm are as follows:

1. Determine the number of clusters
2. Determine the centroid value
   In determining the centroid value for the start of the iteration, the initial centroid value is done randomly. Meanwhile, if determining the centroid value is a stage of iteration, the following formula (1) is used:
   $$\bar{V}_{ij} = \frac{1}{Ni} \sum_{k=0}^{Ni} \times kj \quad (1)$$
3. Calculate the distance between the centroid point and the point of each object using the Euclidean Distance equation (2):
   $$d(ai, bj) = \sqrt{\sum(ai - bj)^2} \quad (2)$$
   d is the distance to the centroid point, ai is the first data attribute and bj is the nth centroid.
4. Grouping objects to determine cluster members by taking into account the minimum distance of objects.
5. Return to step 2, and repeat the process until the value of each centroid stabilizes and the cluster members do not shift to another cluster[14].

3.0 METHODOLOGY

Before proceeding to the data mining phase, operational data must go through the data selection stage. This stage involves selecting relevant and important data attributes from a larger dataset for analysis. In this case, the data selected for this phase includes the historical usage of tractor spare parts from 2023-2024 and the lead time of these spare parts. During this stage, the relevant and important attributes for the k-means clustering data mining process include three attributes: material code, material description, and quantity in a unit of entry, which will then be used in the data mining process. Unnecessary data attributes are removed from the file. Next, using the pivot table feature in Microsoft Excel, the frequency of spare part usage per material is determined. The resulting data is shown in Table 1 below:

Table 1 Data on Frequency of Use of Spare parts

No.	Material	Lead Time	Frequency
1	SFMS0108141	60	9
2	SFMS0108162	60	16
3	SFMS0108231	30	6
4	SFMS0108359	60	5
5	SFMS0108475	60	6
6	SFMS0108597	60	6
7	SFMS0108613	75	1
8	SFMS0108628	60	5
9	SFMS0108651	60	10
10	SFMS0108673	60	6
11	SFMS0108674	60	10
12	SFMS0108685	60	10
13	SFMS0108700	15	12
14	SFMS0108790	60	11
15	SFMS0108791	60	6

No.	Material	Lead Time	Frequency
16	SFMS0108794	60	11
17	SFMS0108824	30	9
...	.....	.....	.....
...	.....	.....	.....
350	SMAT0700214	30	409

Table 1 is the data that will be used in this research. This data has gone through a data mining process in the form of selection and integration. The sample data used was 350 materials.

In this research, the number of clusters that will be used is 5 clusters, this number is the most optimal number of clusters to form a cluster and the proof has been tested using the help of the rapidminer application. Using rapidminer, the average centroid distance is calculated with k=n, then visualized using the elbow method as in Figure 1. The Elbow Method is used to determine the number of good clusters, good cluster results can be used to maximize cluster results[15].

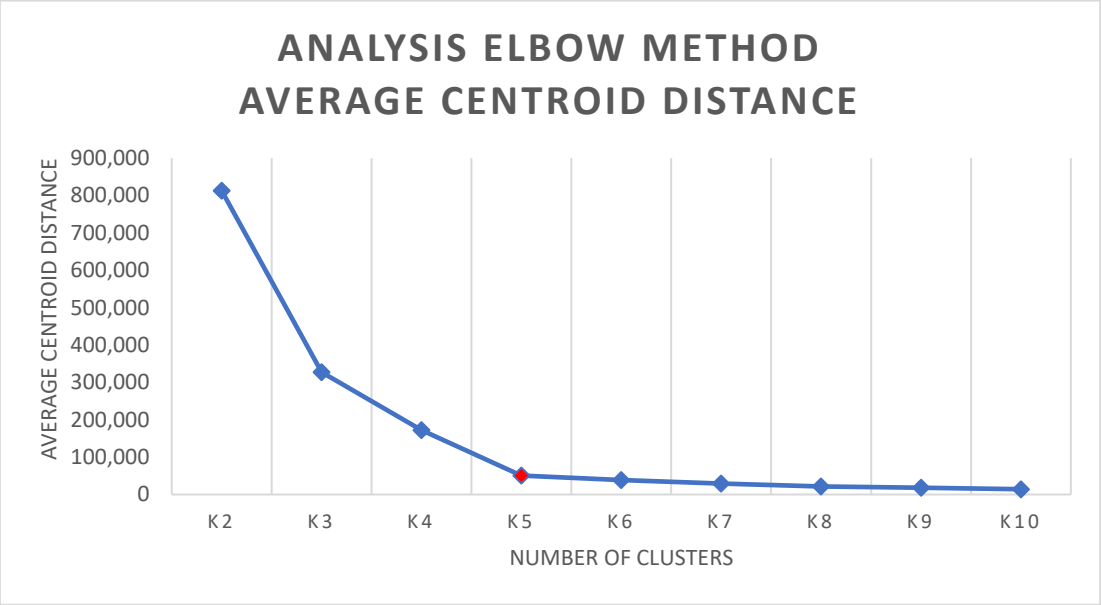

Figure 1 Elbow Method Visualization

1. Number of cluster $k = 5$.
2. The value of k in this study is k = 5, so the number of centroids will be formed as 5 centroids. Determination of the centroid at this initial stage is carried out randomly. The initial centroids selected are the 17th data, 208th data, 18th data, 316th data, and 311th data, namely items SFMS0108824, SSPA0318778, SMAT0700214, SSPA0323748 and SSPA0323707. The initial centroid can be seen in table 2.

Table 2 Initial Centroid Data for Iteration 1

Centroid	Material	Lead Time	Frequency	th Data
C0	SFMS0108824	30	9	17
C1	SSPA0318778	75	3	208
C2	SMAT0700214	30	409	18
C3	SSPA0323748	30	63	316
C4	SSPA0323707	30	207	311

3. The first stage is calculating the distance between the first material data and the first centroid using the equation:

$$d(a_1, b_0) = \sqrt{(60-30)^2 + (9-9)^2} = 30 \quad (1)$$

4. The second stage is calculating the distance between the first material data and the second centroid using the equation:
   $d(a_1, b_1) = \sqrt{(60-75)^2 + (9-3)^2} = 16{,}1$ (2)

5. The third stage is calculating the distance between the first material data and the third centroid using the equation:
   $d(a_1, b_2) = \sqrt{(60-30)^2 + (9-409)^2} = 401$ (3)

6. The fourth stage is calculating the distance between the first material data and the fourth centroid using the equation:
   $d(a_1, b_3) = \sqrt{(60-30)^2 + (9-63)^2} = 61{,}7$ (4)

7. The fifth stage is calculating the distance between the first material data and the fifth centroid using the equation:
   $d(a_1, b_4) = \sqrt{(60-30)^2 + (9-207)^2} = 200{,}2$ (5)

8. The process of calculating the data distance from data 1 to 350 to each centroid in the 1st iteration, assisted by using Microsoft Excel.
9. After obtaining the distance data, the next step is to group the data based on the closest distance to each centroid. The results of the data grouping according to the proximity to each centroid are shown. The results from the first iteration can be seen in Table 3 below.

Table 3 Results of Clustering Iteration 1

No	Distance of Data to Centroid					Min (D1, D2, D3, D4, D5)	Cluster
	C0	C1	C2	C3	C4		
1	30,0000	16,1555	401,1234	61,7738	200,2598	16,1555	1
2	30,8058	19,8494	394,1434	55,7584	193,3417	19,8494	1
3	3,0000	45,0999	403,0000	57,0000	201,0000	3,0000	0
4	30,2655	15,1327	405,1123	65,2993	204,2156	15,1327	1
5	30,1496	15,2971	404,1151	64,4127	203,2265	15,2971	1
6	30,1496	15,2971	404,1151	64,4127	203,2265	15,2971	1
7	45,7056	2,0000	410,4741	76,6094	210,8578	2,0000	1
8	30,2655	15,1327	405,1123	65,2993	204,2156	15,1327	1
9	30,0167	16,5529	400,1262	60,9016	199,2712	16,5529	1
10	30,1496	15,2971	404,1151	64,4127	203,2265	15,2971	1
11	30,0167	16,5529	400,1262	60,9016	199,2712	16,5529	1
12	30,0167	16,5529	400,1262	60,9016	199,2712	16,5529	1
13	15,2971	60,6712	397,2833	53,1601	195,5761	15,2971	0
14	30,0666	17,0000	399,1291	60,0333	198,2826	17,0000	1
15	30,1496	15,2971	404,1151	64,4127	203,2265	15,2971	1
16	30,0666	17,0000	399,1291	60,0333	198,2826	17,0000	1
17	0,0000	45,3982	400,0000	54,0000	198,0000	0,0000	0
18	400,0000	408,4862	0,0000	346,0000	202,0000	0,0000	2
19	433,0000	441,3004	33,0000	379,0000	235,0000	33,0000	2
...	.....	.....	.....	.....	.....	.....	.....
...	.....	.....	.....	.....	.....	.....	.....
350	5,0000	46,3249	395,0000	49,0000	193,0000	5,0000	0

10. Recalculate the new cluster center. Table 4 is the new centroid table after going through calculations on the initial cluster and will be used in the 2nd iteration calculation.

Table 4 New Centroid Data for Iteration 2

Centroid	Lead Time	Frequency
C0	29,875	9,75
C1	73,7019231	3,889423077
C2	30	425,5
C3	31,7647059	63,23529412
C4	30	212,6666667

11. After obtaining the new centroids, recalculate the distances between the material data points and the new centroids using Microsoft Excel. Repeat the calculations until no members of any cluster change their cluster assignment. In this research, for $k = 5$ the iterations were sufficient up to the second iteration. The final results of the calculations for the second iteration can be seen in Table 5 below.

Table 5 Results of Clustering Iteration 1

No	Distance of Data to Centroid					Min (D1, D2, D3, D4, D5)	Cluster
	C0	C1	C2	C3	C4		
1	30,13433	14,62398	417,579	61,1449	205,8643	14,6239766	1
2	30,76651	18,28685	410,5974	55,03094	198,9416	18,2868469	1
3	3,752083	43,75286	419,5	57,26249	206,6667	3,75208275	0
4	30,49718	13,74686	421,5688	64,71925	209,8224	13,746857	1
5	30,35751	13,86352	420,5713	63,82093	208,8327	13,8635216	1
6	30,35751	13,86352	420,5713	63,82093	208,8327	13,8635216	1
7	45,96551	3,167613	426,8785	75,77943	216,3973	3,16761257	1
8	30,49718	13,74686	421,5688	64,71925	209,8224	13,746857	1
9	30,12604	15,00273	416,5816	60,25967	204,875	15,002728	1
10	30,35751	13,86352	420,5713	63,82093	208,8327	13,8635216	1
11	30,12604	15,00273	416,5816	60,25967	204,875	15,002728	1
12	30,12604	15,00273	416,5816	60,25967	204,875	15,002728	1
13	15,04421	59,25958	413,772	53,90835	201,2265	15,0442057	0
14	30,15092	15,43707	415,5842	59,37809	203,8859	15,4370658	1
15	30,35751	13,86352	420,5713	63,82093	208,8327	13,8635216	1
16	30,15092	15,43707	415,5842	59,37809	203,8859	15,4370658	1
17	0,760345	43,99973	416,5	54,264	203,6667	0,76034532	0
18	399,25	407,461	16,5	345,7692	196,3333	16,5	2
19	432,25	440,2848	16,5	378,7688	229,3333	16,5	2
...	.....	.....	.....	.....	.....	.....	.....
...	.....	.....	.....	.....	.....	.....	.....
350	4,251838	44,85624	411,5	49,26691	198,6667	4,25183784	0

12. From the results of manual calculations that have been carried out, the number of clusters obtained is 5 clusters with cluster 0 containing 120 materials, cluster 1 containing 208 materials, cluster 2 containing 2 materials, cluster 3 containing 17 materials, cluster 4 containing 3 materials.
13. After obtaining the results for the members of each cluster, the calculated data was tested using the help of the rapidminer application. Test results using rapidminer can be seen in Figure 2.

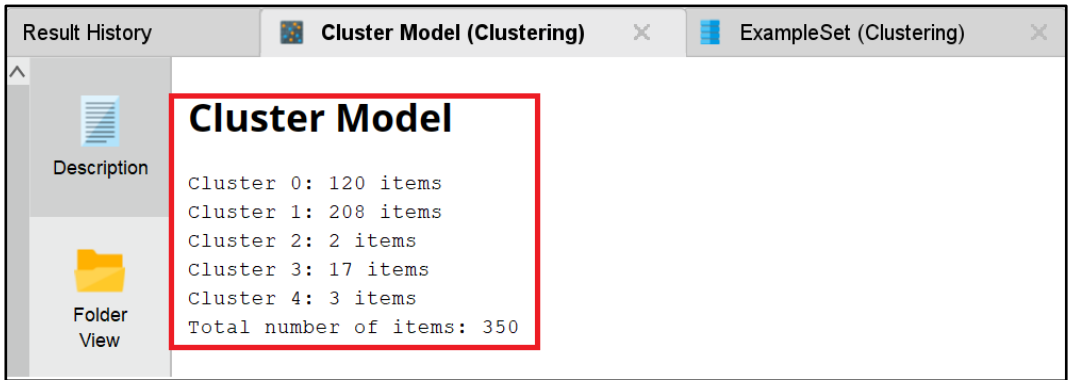

Figure 2 Cluster Test Results $k = 5$ using Rapidminer

14. The test results using rapidminer in Figure 2 are the same as the results of manual calculations using Microsoft Excel. If we visualize the distribution of data for each cluster member, it can be seen in the graph in Figure 3 below.

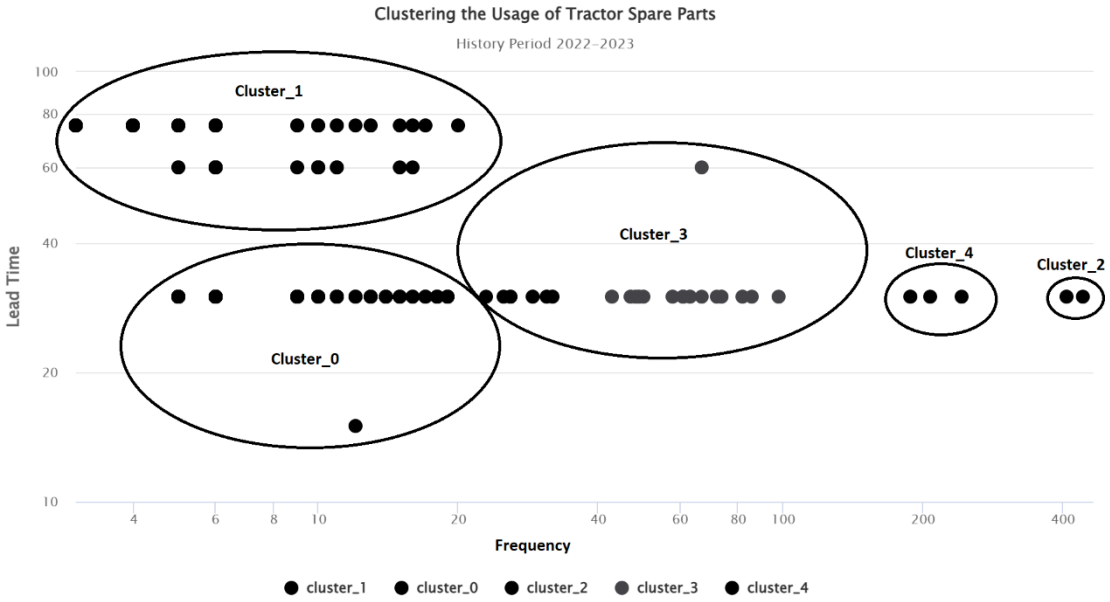

Figure 3 Visualization Results of Cluster Data Distribution

To provide a clear picture of the distribution of data in each cluster, the following is the data and a simple graphical representation in table 6.

Table 6 Cluster Data Representation

Cluster	Number of Cluster	Frequency	Lead Time (days)
0	120	5-40	30
1	208	<20	60-75
2	2	>400	30
3	17	40-100	30
4	3	180-250	30

Cluster 0 : This cluster includes materials that are used with a low frequency, ranging from 5 to 40 times over 2 years. These spare parts are likely used for minor repairs.
Cluster 1 : Materials in this cluster have a low usage frequency, below 20 times over 2 years, and have a longer lead time compared to Cluster 0, ranging from 60 to 75 days.
Cluster 2 : This cluster contains materials with the highest usage frequency compared to those in other clusters, reaching over 400 times within 2 years. Materials in Cluster 2 are very important

and critical for daily operations, so maintaining stock availability in the warehouse and expediting procurement processes are essential.
Cluster 3 : Materials in this cluster have a fairly high usage frequency, ranging from 40 to 100 times. This indicates that these spare parts are important for operations but are used less frequently compared to those in Cluster 4.
Cluster 4 : Materials in this cluster have a relatively high usage frequency with a consistent lead time of 30 days. This suggests that these materials are important and needed regularly, but not as frequently as those in Cluster 2.

4.0 RESULANTS

The conclusion obtained is that the use of k-means clustering can be used to form clusters of data on the use of spare parts needed for tractors. Grouping material clusters based on the history of spare part usage using the k-means algorithm was carried out until the 2nd iteration, where in the 2nd iteration the data members in the cluster members no longer experienced any movement and form 5 clusters. The cluster accuracy results reached 100% after calculating manually using Microsoft Excel and testing using rapidminer. The results of this clustering can be used by the company to determine or develop better inventory management strategies.

5.0 CONCLUSION

K-means clustering can be used to form clusters of data on the use of tractor spare parts. Grouping material clusters based on the history of spare parts use using the k-means algorithm was carried out until the 2nd iteration. From 350 sample data, the results obtained were cluster 0 contains 120 materials, cluster 1 contains 208 materials, cluster 2 contains 2 materials, cluster 3 contains 17 materials, cluster 4 contains 3 materials. The reserarcher advises the company to give special attention to the inventory of materials in clusters 2 and 4, as these clusters consist of materials with high usage frequency that are crucial and essential for operations. An adequate safety stock level is necessary for Clusters 2 and 4. With sufficient safety stock, inventory levels will be maintained, thereby preventing disruptions to operational activities. The researcher suggests that the company should avoid excessive stock storage for materials in cluster 0 and 1, and instead manage procurement based on demand. With low usage frequency, excessive stock storage will increase warehouse storage costs. Additionally, overstocking can lead to other risks, such as aging inventory, reduced product quality, or damage from prolonged storage, which could ultimately harm the company

REFERENCES

[1] N. Salsabila, "CLASSIFICATION OF GOODS USING THE K-MEANS CLUSTERING METHOD IN DETERMINING GOODS STOCK PREDICTIONS (CASE STUDY : UKM MAR ' AH JILBAB KEDIRI) SKRIPSI Oleh : NAJIA SALSABILA," 2019.

[2] W. W. Kristianto and C. Rudianto, "Application of Data Mining in Product Sales Using the K-Means Clustering Method (Case Study of Kakikaki Shoe Store)," no. 5, pp. 90–98, 2022.

[3] Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad, "Application of Data Mining Using the K-Means Algorithm for Clustering Smokers Aged More than 15 Years," *Bull. Inf. Technol.*, vol. 4, no. 4, pp. 501–507, 2023, doi: 10.47065/bit.v4i4.1067.

[4] … Preddy, P. Marpaung, I. Pebrian, and W. Putri, "Application of Data Mining for Population Density Grouping in Deli Serdang Regency Using the K-Means Algorithm," *J. Ilmu Komput. dan Sist. Inf.*, vol. 6, no. 2, pp. 64–70, 2023.

[5] D. M. B. Sitorus, T. Syaputra, and M. Hutasuhut, "Application of Data Mining on Goods Sales Patterns in Cooperatives Using the FP-Growth Algorithm Method," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 3, no. 2, pp. 101–110, 2024.

[6] Vrantika Br Samosir, Agung Mulyo Widodo, Nizirwan Anwar, Binastya Anggara Sekti, and Nixon Erzed, "Outlier Identification Using Data Mining Clustering Techniques for Tracer Study Data Analysis at the Faculty of Computer Science, Esa Unggul University," *IKRA-ITH Inform. J. Komput. dan Inform.*, vol. 8, no. 1, pp. 162–174, 2024, doi: 10.37817/ikraith-informatika.v8i1.3211.

[7] R. S. Wahono, *Data Mining Data mining*, vol. 2, no. January 2013. 2023. [Online]. Available: https://www.cambridge.org/core/product/identifier/CBO9781139058452A007/type/book_part

[8] M. M. K. Neighbor, "Application of data mining to predict sales of best-selling electronic products using the k-nearest neighbor method," 2018.
[9] R. Gustrianda and D. I. Mulyana, "Application of Data Mining in Selection of Superior Products using the K-Means and K-Medoids Algorithm Method," *J. Media Inform. Budidarma*, vol. 6, no. 1, p. 27, 2022, doi: 10.30865/mib.v6i1.3294.
[10] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.
[11] I. N. Abrar, A. Abdullah, and S. Sucipto, "Liver Disease Classification Using the Elbow Method to Determine Optimal K in the K-Nearest Neighbor (K-NN) Algorithm," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 12, no. 2, pp. 218–228, 2023, doi: 10.32736/sisfokom.v12i2.1643.
[12] P. Rani *et al.*, *No Title*الأنا والآخر ودوي زالغرب, vol. 4, no. 1. 2023. doi: 10.1016/j.fcr.2017.06.020.
[13] N. L. W. S. R. Ginantra *et al.*, *Data Mining and Application of Algorithms*. 2021.
[14] J. Eska, M. F. Larasati, P. Studi, and S. Informasi, "Using K-Means Clustering to Group the English Language Skills of Jason English Course Institute Students," *J. Tek. Inform.*, vol. 3, no. 3, 2022.
[15] N. Syahfitri, E. Budianita, A. Nazir, and I. Afrianty, "Product Grouping Based on Inventory Data Using the Elbow and K-Medoid Methods," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 3, pp. 1668–1675, 2023, doi: 10.30865/klik.v4i3.1525.