International Journal Software Engineering and Computer Science (IJSECS)
4 (1), 2024, 210-224
Published Online April 2024 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN : 2776-4869, E-ISSN : 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v4i1.2229.

RESEARCH ARTICLE Open Access

Application of the K-Medoids Algorithm in Clustering PAM Customers Based on Provinces in Indonesia

Nufaisa Almazar *
Informatics Engineering Study Program, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.
Corresponding Email: nufaisaalmazar@gmail.com.

Mesra Betty Yel
Informatics Engineering Study Program, Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, East Jakarta City, Special Capital Region of Jakarta, Indonesia.

Received: February 7, 2024; Accepted: March 25, 2024; Published: April 20, 2024.

Abstract: Clean water is a basic need that is very important for human life. In Indonesia, many clean water companies supply water for community needs. Indonesia's government-owned clean water company has many customers spread across various provinces. This research aims to apply the k-medoids algorithm in clustering PAM customers based on provinces in Indonesia to determine which provinces are included in the low, medium, and high-level clusters of PAM customers. The data used in this research is on the number of PAM customers, the amount of water distributed, and the value of the water distributed from 1998 to 2021 to 34 provinces in Indonesia. This research obtained 3 clusters, namely low, medium, and high-level clusters determined by the number of PAM customers, the amount of water distributed, and the value of the water distributed in each province. Cluster 1 (C1) is a low-level cluster that obtained results from 19 provinces, Cluster 2 (C2) is a medium-level cluster that obtained results from 11 provinces, and Cluster 3 (C3) is a high-level cluster that obtained results from 4 provinces. And get a k-medoids cluster validation value of 0.014846718. In this way, it is hoped that the government can identify areas in each province that require improvement or improvement in making appropriate decisions regarding clean water resources to provide better services to customers.

Keywords: K-Medoids Algorithm; PAM Customer Clusterization; PAM Water Services.

1. Introduction

Clean water is a vital necessity for human life. Several clean water companies in Indonesia ensure that water supplies meet community needs. The Clean Water Company is responsible for procuring, purifying, providing, and distributing clean water to customers, including households, industry, and other consumers, for commercial purposes. This type of business includes Drinking Water Companies (PAM), Regional Drinking Water Companies (PDAM), and similar entities. Indonesia's government-owned clean water company has several customers spread across various provinces. According to data from the Central Statistics Agency, from 1998 to 2021, PAM customers increased significantly from 43,948 in 1998 to 15,973,088 in 2021. Meanwhile, the volume of water distributed increased from 1,479,953 m3 in 1998 to 4,375—694 m3 in 2021.

The value of the water distributed has also increased from IDR 1,170,652 in 1998 to IDR 18,276,246 in 2021. In managing this significant number of customers, the government must implement effective strategies to meet their needs well. One approach that can be applied is using the k-medoids algorithm to classify PAM customers in all provinces of Indonesia. The k-medoids algorithm is a data clustering analysis method that uses objects in a data set to represent a cluster, with selected objects referred to as medoids. In this research, the data analyzed is the number of PAM customers by province in Indonesia. The application of the k-medoids algorithm has the potential to help companies manage and distribute clean water supplies more efficiently. The government can identify areas with different water needs by classifying PAM customers by province. This allows the government to regulate water distribution more accurately, avoiding problems of shortage or oversupply. Drinking Water Companies (PAM) in Indonesia are very important in providing clean water for the community. As a basic need, clean water is a human right that must be fulfilled by the government and companies involved in providing water services. However, a deep understanding of customer characteristics in various provinces is crucial to ensure that this service can be provided optimally. This not only allows for improving the quality of services but also helps in devising effective business strategies.

This research aims to apply the K-Medoids algorithm to group PAM customers based on provinces in Indonesia. The K-Medoids algorithm was chosen because it has several advantages compared to other clustering algorithms, especially K-Means. One of the main advantages of K-medoids is their robustness to outlier data, which makes them more stable in producing accurate clustering. Apart from that, K-Medoids are also easier to interpret because they use actual data points as medoids, making it easier to understand the characteristics of each cluster. Grouping PAM customers by province has significant benefits for drinking water companies. First, it can help improve service quality by enabling PAM to identify specific needs in each region. Thus, more targeted and effective strategies can be designed to meet customer needs in each province. In addition, information about customer characteristics can also be used to develop more effective business strategies, such as developing new products and services, setting appropriate prices, and optimizing water distribution. With a better understanding of customer needs, PAM can allocate resources more efficiently and increase overall customer satisfaction.

Previous studies have tried to apply clustering algorithms for grouping data in various fields, including sales, customer satisfaction, and determining new customers. However, it is hoped that this research can make a new contribution to grouping PAM customers based on provinces in Indonesia. Thus, the results of this research can provide valuable insight for PAM in improving the quality of its services and the effectiveness of its business strategies. Several previous studies have also studied clustering algorithms in different contexts. Simanjuntak *et al*. (2023) applied the K-Medoids algorithm for grouping unemployment in North Sumatra [1], while Anggarwati *et al*. (2021) used the K-Means algorithm to predict car body sales. Other research, such as that conducted by Pangestu and Ridwan (2022), Simanjuntak (2022), and Fajriana (2021), has also produced valuable insights into the use of clustering algorithms in various business and social [2][3][4][5]. However, although there has been a lot of previous research on grouping data using clustering algorithms, using the K-Medoids algorithm in the context of grouping PAM customers based on provinces in Indonesia still needs to be improved. Therefore, this study aims to fill this gap by investigating the potential and benefits of using the K-Medoids algorithm in this context. Thus, this research can provide a new contribution to our understanding of how clustering algorithms, especially K-Medoids, can be applied to improve the efficiency and effectiveness of clean water services in Indonesia. By understanding the characteristics of PAM customers in various provinces, PAM can develop more targeted and practical strategies to meet community needs better. In addition, this research can also be a basis for further research in this field and inspire similar research in other countries that face similar challenges in providing clean water services to their communities.

2. Research Method

2.1 Research data

This research data uses secondary data collection methods. Secondary data is a source of data that is not directly obtained from objects through interviews. For this research, the data source is a public dataset from the Central Statistics Agency website, linked to https://www.bps.go.id/id. The dataset is data on the number of PAM customers, the amount of water distributed, and the value of the water distributed from 1998 - 2021. This research also uses a literature review of 20 scientific journals as a research reference.

2.2 Research Tools

The research tools used in this research consist of hardware and software. The hardware includes a computer with standard specifications, including a powerful processor and sufficient memory to run the required software. Meanwhile, the software used is Microsoft Excel, version 2013. This software has several primary functions in this research. First, Excel is used to store and manage datasets used in analysis. This dataset includes raw data obtained from sources relevant to the research topic. Excel provides flexibility in organizing and managing this data according to research needs. Besides that, Excel is also used to calculate analysis results manually using the k-medoids algorithm. This algorithm is one of the methods used in data clustering analysis to group data into several groups based on the similarity of specific characteristics. Using Excel, researchers can manually implement this algorithm by referring to the basic principles of k-medoids and applying them to existing datasets. As the leading software in this research, Excel was chosen based on its availability, ease of use, and flexibility in managing data and carrying out fundamental statistical analysis. Even though it is not a special software for data analysis, Excel remains a popular choice among researchers because its capabilities are sufficient for simple to medium analysis purposes.

2.3 Application of Methodology

The methodology for applying the k-means algorithm in clustering customers of clean water companies in all provinces of Indonesia can be carried out in the following steps.

1) Data Collection
   Collect data on the number of PAM customers, the amount of water distributed, and data on the value of the water distributed, including the province where they live.
2) Standardizing Data
   Data standardization is done to reduce data variations between variables because each variable has different units. This will influence the results of analytical decisions because unit differences create an extensive range of values between variables. To calculate data standardization, you can use the following formula.

$$Z = \frac{x_i - \bar{x}}{S_x}$$

Information :
Z = z-score value
xi = value of the ith observation
$\bar{x}$ = mean value
Sx = standard deviationc.

3) Conduct descriptive analysis
   Descriptive analysis is a procedure and method for collecting, presenting, summarizing, and analyzing data. The main goal of descriptive statistics is to collect, present, and summarize a set of data so that it can provide helpful information. Descriptive statistics only concerns the data at hand and does not draw more or further conclusions from the existing data. Examining the essential properties of existing data is called data analysis (description). Descriptive analysis determines the values of max, min, mean, and standard deviation.

4) Carrying out Cluster Analysis
   Cluster analysis is a statistical technique that aims to group objects into groups such that objects in one group will have high similarities compared to objects in another group. There are two assumptions from cluster analysis.

a) Representative Sample
The sample taken can truly represent an existing population or part of a population that attempts to accurately reflect the characteristics of a larger group.

b) Multicollinearity Test
Multicollinearity is the occurrence of a linear relationship between independent variables. A case of multicollinearity is the occurrence of correlation between independent variables, meaning that there is a correlation between $Y1$, $Y2$, $Y3$, ... If there is multicollinearity between grouping variables, note that the set of grouping variables is assumed to be independent, but can actually be correlated. This may be a problem if some variables in a cluster variable set are highly correlated and others are relatively uncorrelated. In such a situation, correlated variables influence the cluster solution more than several uncorrelated variables. The process of determining whether multicollinearity exists or not can be done in the following way:
(1) The tolerance value is the level of error that is justified statistically.
(2) The variance inflation factor (VIF) value is the squared standard deviation inflation factor.

Calculation of the VIF value can use the following formula:

$$VIF\ (i) = \frac{1}{1-\mathrm{Ri2}}$$

Where:
i = Several of independent variables, i:1,2,3.....n.
$R_1{}^2$ = Coefficient of determination.

As for finding value $R_1{}^2$ can use the following formula:

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

RSS = Sum of residual squares
TSS = Sum of total squares
$R_1{}^2$ = Coefficient of determination.

If the VIF value is less than 10 then multicollinearity does not occur.

5) Determining the Number of Clusters
Determine the number of clusters that will be formed based on the number of provinces in Indonesia. For example, if there are 34 provinces in Indonesia, then the desired number of clusters can be determined, namely 3 clusters.
6) Initializing Centroid
Initialize the cluster center point randomly (randomly) as initial centroid initialization for each cluster.
7) K-medoids Iteration
Iterate the k-medoids method to group customer data into clusters based on province. This iteration involves calculating the distance between calculating the distance between each customer and each cluster centroid using the Euclidean Distance method. Formula:

$$\mathrm{d\ (p,q)} = \sqrt{((p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2)}$$

8) Grouping data
Group data into clusters that have the closest distance to the data. The data that has the smallest distance to the cluster centroid will become a member of the cluster.
9) Repeating steps e – g
Repeat steps e – g until there is no further change in the position of the cluster center or the iteration has reached a previously determined limit.
10) Calculate the total deviation (S) by calculating the value of the new total distance – the old total distance. If the result S < 0, then exchange the objects with the cluster data to form a new set of k objects as medoids. If the result S > 0, then the iterative calculation process is stopped.

11) Validating k-medoids clusters
Clustering validity is carried out to measure how well and accurately the clusters are formed. Cluster validation is carried out by calculating the cluster variance value. The smaller the variance value between members, the better the results will be, and the greater the variance value between clusters, the better the results obtained. Validate the variance values within clusters (Vw) and variance between clusters (Vb). The following is the formula for

$$V_w: \qquad V_w = \frac{1}{\mathrm{N-k}}\sum^{k}_{i=1}(n_i - 1) \times V_i^2$$

Where:
V_w = Variance within clusters
k = Many clusters
ni = The amount of data in a cluster
N = i-th data in a cluster
V_i^2 = Variance in a cluster-i

The following is the formula

$$V_b: \qquad V_b = \frac{1}{\mathrm{k-1}}\ \sum^{k}{}_{i=1}\ n_i\ (\bar{x_i} - \bar{x})^2$$

Where:
V_b = Variance between clusters
k = Many clusters
ni = The amount of data in a cluster
$\bar{x}i$ = The average of the i-th data in a cluster
$\bar{x}$ = The average of all data

The following is the formula for finding the variance of the entire cluster (V):

$$V = \frac{\mathrm{Vw}}{\mathrm{Vb}}$$

Where:
V = Variance
V_w = Variance within clusters
V_b = Variance between clusters.

3. Result and Discussion

3.1 Results

3.1.1 Calculating Manual Data Using Microsoft Excel
In this research, data on the number of PAM customers, the amount of water distributed, and the value of the water distributed can be seen in Table 1. This research uses Microsoft Excel to calculate the data manually.

Table 1. PAM Customer Data Used

No.	Province	PAM Customers	Distributed Water	Value of Distributed Water
1	Aceh	1019686	538590	1215183
2	Sumatera Utara	4383302	4997074	11007433
3	Sumatera Barat	2895980	1306158	3240371
4	Riau	448626	494768	1434623
5	Jambi	4389332	630443	1694807
6	Sumatera Selatan	4280916	2030799	6244270
7	Bengkulu	451973	306710	734713
8	Lampung	433837	382192	1016162
9	Kep. Bangka Belitung	155760	94272	328785
10	Kep. Riau	1680818	1196203	6227481
11	Dki Jakarta	4511107	10213953	50498748
12	Jawa Barat	8900810	5752421	22055776

13	Jawa Tengah	8596373	6084046	15655577
14	DI Yogyakarta	926033	617121	1893693
15	Jawa Timur	10751827	11121559	23321542
16	Banten	1638449	3022090	10735930
17	Bali	2362335	2357960	7259249
18	Nusa Tenggara Barat	2646576	1017669	2009997
19	Nusa Tenggara Timur	781581	514204	1274309
20	Kalimantan Barat	1255585	932172	2781799
21	Kalimantan Tengah	755588	494704	1577441
22	Kalimantan Selatan	2348471	1394604	5657187
23	Kalimantan Timur	2421851	2247148	8637958
24	Kalimantan Utara	338293	148385	658621
25	Sulawesi Utara	600864	422721	1371996
26	Sulawesi Tengah	511177	348530	597615
27	Sulawesi Selatan	2541939	1853189	5808079
28	Sulawesi Tenggara	475572	248580	882239
29	Gorontalo	408508	240152	642633
30	Sulawesi Barat	295623	103156	308630
31	Maluku	290489	170962	587296
32	Maluku Utara	424823	229992	700423
33	Papua Barat	127689	65074	297154
34	Papua	376318	368443	1058061

Information:

a. Data on the number of PAM customers was obtained from the total number of clean water customers from 1998-2021.
b. Data on the amount of water distributed was obtained from the total sum of water distributed from 1998-2021.
c. Data on the value of water distributed is obtained from the total sum of the values of water distributed from 1998-2021.

The following table 2 shows the results of standard deviation calculations. Variable X1 obtained a standard deviation of 2663951.589; Variable X2 obtained a standard deviation of 2737294.774, and Variable.

Table 2. Standard Deviation Calculation Results

Variable	N	Min	Max	Mean	Standard Deviation
X1	34	74428111	10751827	2189062	2663951,589
X2	34	61946044	11121559	1821942	2737294,774
X3	34	199415781	50498748	5865170	9858509,814

Information

a. PAM Customer: Variable X1
b. Distributed Water: Variable X2
c. Value of Water Distributed: Variable X3

The following are Tables 3 and 4, which show the results of data standardization calculations and the results of standard deviation calculations. Variable X1 gets a standard deviation of 1, Variable X2 gets a standard deviation of 1, and Variable X3 gets a standard deviation of 1.

Table 3. Data Standardization Calculation Results

No.	Province	X1	X2	X3
1	Aceh	-0,438962965	-0,468839704	-0,471672405
2	Sumatera Utara	0,823678599	1,159952359	0,521606517
3	Sumatera Barat	0,265364399	-0,188428545	-0,266247037

4	Riau	-0,653328722	-0,48484894	-0,449413462
5	Jambi	0,825942153	-0,435283581	-0,423021644
6	Sumatera Selatan	0,785244717	0,076300343	0,038454085
7	Bengkulu	-0,652072318	-0,553551077	-0,520408979
8	Lampung	-0,65888025	-0,525975677	-0,491860141
9	Kep. Bangka Belitung	-0,763265405	-0,631159818	-0,561584371
10	Kep. Riau	-0,190785782	-0,228597766	0,036751089
11	Dki Jakarta	0,87165432	3,065804461	4,527416294
12	Jawa Barat	2,519470676	1,435898891	1,642297495
13	Jawa Tengah	2,405190447	1,557049526	0,993091974
14	DI Yogyakarta	-0,474118634	-0,44015043	-0,402847601
15	Jawa Timur	3,214309505	3,39737489	1,770690733
16	Banten	-0,206690351	0,438442926	0,494066554
17	Bali	0,065043566	0,195820171	0,141408691
18	Nusa Tenggara Barat	0,171742577	-0,293820555	-0,391050281
19	Nusa Tenggara Timur	-0,528343343	-0,477748499	-0,465674946
20	Kalimantan Barat	-0,35041068	-0,325054678	-0,312762384
21	Kalimantan Tengah	-0,538100653	-0,484872321	-0,434926689
22	Kalimantan Selatan	0,059839268	-0,156117081	-0,021096802
23	Kalimantan Timur	0,087384813	0,155337866	0,281258326
24	Kalimantan Utara	-0,694745766	-0,61139103	-0,528127387
25	Sulawesi Utara	-0,596181288	-0,511169452	-0,455766045
26	Sulawesi Tengah	-0,629848191	-0,53827322	-0,534315544
27	Sulawesi Selatan	0,13246371	0,011415113	-0,00579104
28	Sulawesi Tenggara	-0,643213674	-0,574787372	-0,505444649
29	Gorontalo	-0,668388305	-0,577866325	-0,529749133
30	Sulawesi Barat	-0,710763325	-0,627914278	-0,563628797
31	Maluku	-0,712690537	-0,603143106	-0,535362253
32	Maluku Utara	-0,662263945	-0,581578019	-0,523887193
33	Papua Barat	-0,773802759	-0,641826554	-0,564792868
34	Papua	-0,680471858	-0,530998519	-0,487610107

Table 4. Standard Deviation After Data Standardization

Variable	N	Min	Max	Mean	Standard Deviation
X1	34	-0,773802759	3,214309505	0	1
X2	34	-0,641826554	3,39737489	0	1
X3	34	-0,564792868	4,527416294	0	1

Tables 5 and 6 show the results of calculating correlation values, Tolerance values, and Variance Inflation Factor (VIF). The rx1x2 results got a tolerance value of 0.281431091 and a VIF value of 3.553267676, the rx1x2 results got a tolerance value of 0.556156536 and a VIF value of 1.798054926, and the rx2x3 results got a tolerance value of 0.166044915 and a VIF value of 6.022466862. If the calculation results show a tolerance value > 0.10 and a VIF value < 10, then multicollinearity does not occur.

Table 5. Correlation Results

	X1	X2	X3
X1	1		
X2	0,847684439	1	
X3	0,666215779	0,913211413	1

Table 6. Calculation Results of Tolerance and Variance Inflation Factor (VIF) Values

	r	r^2	Tolerance	VIF
rx1x2	0,847684439	0,718568909	0,281431091	3,553267676
rx1x3	0,666215779	0,443843464	0,556156536	1,798054926
rx2x3	0,913211413	0,833955085	0,166044915	6,022466862

Following are the steps to calculate the k-medoids algorithm manually:

a. Determine the desired number of clusters (k), namely 3 clusters.
b. Choose randomly to determine the initial medoid as in table 7.

Table 7. Initial Medoids

Clusters	PAM Customers	Distributed Water	Value of Distributed Water
C1	-0,668388305	-0,577866325	-0,529749133
C2	0,171742577	-0,293820555	-0,391050281
C3	0,785244717	0,076300343	0,038454085

Table 7 is the basis for determining the clusters in the initial medoids, namely cluster 1 (C1) is a low-level cluster with a low number of PAM customers, the amount of water distributed, and the value of the water distributed, cluster 2 (C2) is a medium level cluster with a large number of PAM customers, the amount of water distributed. The value of the water distributed is moderate, and cluster 3 (C3) is a high-level cluster with a high number of PAM customers, the amount of water distributed, and the value of the water distributed. Calculate the distance of the object to each selected initial medoid using the distance measurement equation Euclidean Distance.

$$d\ (p, q)$$

The following is a sample calculation of the distances C1, C2, and C3 in Aceh province.

dAceh, C1 = √(-0,438962965-(-0,668388305))2 + (-0,468839704-(-0,577866325))2 + (-0,471672405-(-0,529749133))2)
= 0
dAceh, C2 = √(-0,438962965-0,171742577)2 + (-0,468839704-0,293820555)2 + (-0,471672405-(-0,391050281))2)
= 1
dAceh, C3 = √(-0,438962965-0,785244717)2 + (-0,468839704-0,076300343)2 + (-0,471672405-0,038454085)2)
= 1

Determining the closest distance in iteration 1 is taken from the results of the smallest distance value. Distance calculations for C1, C2, and C3 can be seen in table 8.

Table 8. Results of Distance Values (Cost) in Iteration 1

Provinsi	C1	C2	C3	Jarak Terdekat	Hasil Cluster
Aceh	0	1	1	0	1
Sumatera Utara	3	2	1	1	3
Sumatera Barat	1	0	1	0	2
Riau	0	1	2	0	1
Jambi	2	1	1	1	2
Sumatera Selatan	2	1	0	0	3
Bengkulu	0	1	2	0	1
Lampung	0	1	2	0	1
Kep. Bangka Belitung	0	1	2	0	1
Kep. Riau	1	1	1	1	2
Dki Jakarta	6	6	5	5	3
Jawa Barat	4	4	3	3	3
Jawa Tengah	4	3	2	2	3
DI Yogyakarta	0	1	1	0	1
Jawa Timur	6	5	4	4	3
Banten	2	1	1	1	3
Bali	1	1	1	1	2
Nusa Tenggara Barat	1	0	1	0	2
Nusa Tenggara Timur	0	1	2	0	1
Kalimantan Barat	0	1	1	0	1
Kalimantan Tengah	0	1	2	0	1
Kalimantan Selatan	1	0	1	0	2
Kalimantan Timur	1	1	1	1	3
Kalimantan Utara	0	1	2	0	1
Sulawesi Utara	0	1	2	0	1
Sulawesi Tengah	0	1	2	0	1
Sulawesi Selatan	1	0	1	0	2
Sulawesi Tenggara	0	1	2	0	1
Gorontalo	0	1	2	0	1
Sulawesi Barat	0	1	2	0	1
Maluku	0	1	2	0	1
Maluku Utara	0	1	2	0	1
Papua Barat	0	1	2	0	1
Papua	0	1	2	0	1
Jumlah	38	42	54		
Total Cost	134				

C1	C2	C3
1		
		1
	1	
1		
	1	
		1
1		
1		
1		
	1	
		1
		1
		1
1		
		1
		1
	1	
	1	
1		
1		
1		
	1	
		1
1		
1		
1		
	1	
1		
1		
1		
1		
1		
1		

Total Hasil Cluster

Cluster		
1	2	3
19	7	8

Table 8 shows the total results of 3 temporary clusters, namely:

1) Cluster 1 (C1) is a high-level cluster with 19 provinces, namely Aceh, Riau, Bengkulu, Lampung, Kep. Bangka Belitung, DI Yogyakarta, East Nusa Tenggara, West Kalimantan, Central Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua and Papua.
2) Cluster 2 (C2) is a medium level cluster with 7 provinces, namely West Sumatra, Jambi, Kep. Riau, Bali, West Nusa Tenggara, South Kalimantan and South Sulawesi.
3) Cluster 3 (C3) is a low level cluster with 8 provinces, namely: North Sumatra, South Sumatra, DKI Jakarta, West Java, Central Java, East Java, Banten and East Kalimantan.

Choose randomly to determine the 2nd medoid as in table 9.

Table 9. 2nd Medoids

Clusters	PAM Customers	Distributed Water	Value of Distributed Water
C1	-0,629848191	-0,53827322	-0,534315544
C2	0,059839268	-0,156117081	-0,021096802
C3	2,519470676	1,435898891	1,642297495

Table 9 is the basis for determining clusters in the 2nd medoids, namely cluster 1 (C1) is a low-level cluster with the number of PAM customers, the amount of water distributed, and the value of the water distributed in low amounts, cluster 2 (C2) is a medium level cluster with a moderate number of PAM customers, the amount of water distributed, and the value of the water distributed. Cluster 3 (C3) is a high-level cluster with many PAM customers, the amount of water distributed, and the value of the water distributed. The method for calculating the object's distance to each selected second medoid is the same as the initial medoids using the distance measurement equation Euclidean Distance. The following is a sample calculation of the distances C1, C2, and C3 in Aceh province.

dAceh, C1 = √(-0,438962965-(-0,629848191))2 + (-0,468839704-(-0,53827322))2 + (-0,471672405-(-0,534315544))2)

= 0

dAceh, C2 = √(-0,438962965-(0,059839268))2 + (-0,468839704-(-0,156117081))2 + (-0,471672405-(-0,021096802))2)

= 1

dAceh, C3 = √(-0,438962965-2,519470676)2 + (-0,468839704-1,435898891)2 + (-0,471672405-1,642297495)2)
= 4

The method for calculating the object's distance to each selected second medoid is the same as the initial medoids using the distance measurement equation Euclidean Distance. The following is a sample calculation of the distances C1, C2, and C3 in Aceh province.

Table 10. Results of Distance Values (Cost) in Iteration 2

Provinsi	C1	C2	C3	Jarak Terdekat	Hasil Cluster	C1	C2	C3
Aceh	0	1	4	0	1	1		
Sumatera Utara	2	2	2	2	2		1	
Sumatera Barat	1	0	3	0	2		1	
Riau	0	1	4	0	1	1		
Jambi	1	1	3	1	2		1	
Sumatera Selatan	2	1	3	1	2		1	
Bengkulu	0	1	4	0	1	1		
Lampung	0	1	4	0	1	1		
Kep. Bangka Belitung	0	1	4	0	1	1		
Kep. Riau	1	0	4	0	2		1	
Dki Jakarta	6	6	4	4	3			1
Jawa Barat	4	3	0	0	3			1
Jawa Tengah	4	3	1	1	3			1
DI Yogyakarta	0	1	4	0	1		1	
Jawa Timur	6	5	2	2	3			1
Banten	1	1	3	1	2		1	
Bali	1	0	3	0	2		1	
Nusa Tenggara Barat	1	0	4	0	2		1	
Nusa Tenggara Timur	0	1	4	0	1	1		
Kalimantan Barat	0	1	4	0	1	1		
Kalimantan Tengah	0	1	4	0	1	1		
Kalimantan Selatan	1	0	3	0	2		1	
Kalimantan Timur	1	0	3	0	2		1	
Kalimantan Utara	0	1	4	0	1	1		
Sulawesi Utara	0	1	4	0	1	1		
Sulawesi Tengah	0	1	4	0	1	1		
Sulawesi Selatan	1	0	3	0	2		1	
Sulawesi Tenggara	0	1	4	0	1	1		
Gorontalo	0	1	4	0	1	1		
Sulawesi Barat	0	1	4	0	1	1		
Maluku	0	1	4	0	1	1		
Maluku Utara	0	1	4	0	1	1		
Papua Barat	0	1	4	0	1	1		
Papua	0	1	4	0	1	1		
Jumlah	37	41	122					
Total Cost	200							

Total Hasil Cluster

Cluster		
1	2	3
19	11	4

Table 10 shows the total results of 3 clusters, namely:

1) Cluster 1 (C1) is a high-level cluster with 19 provinces, namely Aceh, Riau, Bengkulu, Lampung, Kep. Bangka Belitung, DI Yogyakarta, East Nusa Tenggara, West Kalimantan, Central Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua and Papua.
2) Cluster 2 (C2) is a medium level cluster with 11 provinces, namely North Sumatra, West Sumatra, Jambi, South Sumatra, Kep. Riau, Banten, Bali, West Nusa Tenggara, South Kalimantan, East Kalimantan and South Sulawesi.
3) Cluster 3 (C3) is a low level cluster containing 4 provinces, namely: DKI Jakarta, West Java, Central Java and East Java Provinces.

Calculating Total Deviation (S): Calculate the total deviation (S) by calculating the value of the new total distance – the old total distance. If the result S < 0, then exchange the objects with the cluster data to form a new set of k objects as medoids. If the result S > 0, then the iterative calculation process is stopped.

S = New Total Cost – Old Total Cost
= 200 - 134
= 66

3.1.2 Data Calculation Results Using Microsoft Excel

This research uses a manual calculation process using Microsoft Excel on PAM customer data based on provinces in Indonesia from 1998-2021. The results of these calculations are in tables 11, 12, and 13.

Table 11. Manual calculation results including Cluster 1

No.	Province	X1	X2	X3
1	Aceh	-0,438962965	-0,468839704	-0,471672405
2	Riau	-0,653328722	-0,48484894	-0,449413462
3	Bengkulu	-0,652072318	-0,553551077	-0,520408979
4	Lampung	-0,65888025	-0,525975677	-0,491860141
5	Kep. Bangka Belitung	-0,763265405	-0,631159818	-0,561584371
6	DI Yogyakarta	-0,474118634	-0,44015043	-0,402847601
7	Nusa Tenggara Timur	-0,528343343	-0,477748499	-0,465674946
8	Kalimantan Barat	-0,35041068	-0,325054678	-0,312762384
9	Kalimantan Tengah	-0,538100653	-0,484872321	-0,434926689
10	Kalimantan Utara	-0,694745766	-0,61139103	-0,528127387
11	Sulawesi Utara	-0,596181288	-0,511169452	-0,455766045
12	Sulawesi Tengah	-0,629848191	-0,53827322	-0,534315544
13	Sulawesi Tenggara	-0,643213674	-0,574787372	-0,505444649
14	Gorontalo	-0,668388305	-0,577866325	-0,529749133
15	Sulawesi Barat	-0,710763325	-0,627914278	-0,563628797
16	Maluku	-0,712690537	-0,603143106	-0,535362253
17	Maluku Utara	-0,662263945	-0,581578019	-0,523887193
18	Papua Barat	-0,773802759	-0,641826554	-0,564792868
19	Papua	-0,680471858	-0,530998519	-0,487610107

Table 12. Manual calculation results including Cluster 2

No.	Province	X1	X2	X3
1	Sumatera Utara	0,823678599	1,159952359	0,521606517
2	Sumatera Barat	0,265364399	-0,188428545	-0,266247037
3	Jambi	0,825942153	-0,435283581	-0,423021644
4	Sumatera Selatan	0,13246371	0,011415113	-0,00579104
5	Kep. Riau	-0,190785782	-0,228597766	0,036751089
6	Banten	-0,206690351	0,438442926	0,494066554
7	Bali	0,065043566	0,195820171	0,141408691
8	Nusa Tenggara Barat	0,171742577	-0,293820555	-0,391050281
9	Kalimantan Selatan	0,059839268	-0,156117081	-0,021096802
10	Kalimantan Timur	0,087384813	0,155337866	0,281258326
11	Sulawesi Selatan	0,785244717	0,076300343	0,038454085

Table 13. Manual calculation results including Cluster 3

No.	Province	X1	X2	X3
1	Dki Jakarta	0,87165432	3,065804461	4,527416294
2	Jawa Barat	2,519470676	1,435898891	1,642297495
3	Jawa Tengah	2,405190447	1,557049526	0,993091974
4	Jawa Timur	3,214309505	3,39737489	1,770690733

The basis for determining clusters is that cluster 1 (C1) is a low-level cluster because the number of PAM customers, the amount of water distributed, and the value of the water distributed are low, cluster 2 (C2) is a medium-level cluster because of the number of PAM customers, the amount of water distributed, and the value of the water distributed is moderate. Cluster 3 (C3) is a high-level cluster because the number of PAM

customers, the amount of water distributed, and the value of the water distributed are high. Tables 11, 12, and 13 show the total results of 3 clusters, namely.

a. Cluster 1 (C1) is a low level cluster with 21 provinces, namely Aceh, Riau, Jambi, Bengkulu, Lampung, Kep. Bangka Belitung, DI Yogyakarta, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua and Papua.
b. Cluster 2 (C2) is a medium level cluster containing 9 provinces, namely North Sumatra, West Sumatra, South Sumatra, Kep. Riau, Banten, Bali, South Kalimantan, East Kalimantan and South Sulawesi.
c. Cluster 3 (C3) is a low level cluster consisting of 4 provinces, namely: DKI Jakarta, West Java, Central Java and East Java Provinces.

The following is a table of 14 k-medoids cluster validation results on PAM customer data based on provinces in Indonesia from 1998 to 2021. There is a lot of data for 34 provinces in Indonesia with 3 clusters. The variance value in the cluster (Vw) of variables X1, between clusters (Vb) from X1, and variable X3 gets a result of 0.022177788. The k-medoids cluster validation value obtained a result of 0.014846718, which was obtained from the k-medoids VAR value. It can be said that the cluster has good variance because it is close to 0, and the smaller the cluster variance, the better.

Table 14. K-Medoids Cluster Validation Results

N(banyaknya data)	34	34	34
k(banyaknya cluster)	3	3	3
Vw	0	0	0
Vb	14	14	12
Vw/Vb	0,010486932	0,011875432	0,022177788
Var K Medoids	0,014846718		

K-Medoids	**X1**	**X2**	**X3**
Vw	0	0	0
Vb	14	14	12
Vw/Vb	0,010486932	0,011875432	0,022177788
Var K Medoids	0,014846718		

Information :

a. Vw : Within-cluster variance
b. Vb: Variance between clusters
c. Vw/Vb: is the formula used to find the variance value of the entire cluster (V)
d. VAR K-medoids: Calculation results are obtained from the average values of variables X1, X2 and X3.

3.2 Discussion

Analysis carried out on PAM customer data based on provinces in Indonesia from 1998 to 2021 provides an in-depth understanding of water use distribution patterns and characteristics in various regions. The steps taken in processing this data involve using Microsoft Excel to carry out manual calculations and applying the k-medoids algorithm for cluster division. First, data on the number of PAM customers, the amount of water distributed, and the value of the water distributed are collected and entered into a table. Microsoft Excel makes it possible to carry out manual calculations accurately and efficiently. The table is then used to calculate standard deviation, data standardization, correlation values, and the distance between objects using Euclidean Distance. The analysis results show that the data can be divided into three main clusters, namely low, medium, and high-level clusters, based on water use characteristics. These clusters show different patterns regarding the number of PAM customers, the amount of water distributed, and the value of the water distributed in each region. Low-level clusters tend to have lower numbers of customers and water volumes, while high-level clusters have higher numbers in both aspects. The k-medoids algorithm divides data into clusters by randomly selecting initial medoids and calculating the distance between objects and medoids using Euclidean Distance. After several iterations, the final clusters were determined, and the results showed a diverse distribution of water use in various provinces in Indonesia. Next, cluster validation was done using the Variance Ratio (VAR)

k-medoids. The validation results show that the clusters have good variance because it is close to zero, which indicates a high level of homogeneity in terms of water use in each cluster. This shows that cluster division with the k-medoids algorithm provides consistent and reliable results. This analysis provides a deep understanding of water use distribution patterns in Indonesia and allows for better decision-making in water resource planning and management. This information can be used by governments, water companies, and other related organizations to develop more effective strategies to support sustainable water use in various regions. Apart from that, this analysis also provides an overview of the challenges and opportunities faced in water resources management in Indonesia. With a better understanding of water use patterns, stakeholders can identify areas that require special attention and develop appropriate solutions to increase the efficiency and sustainability of water use.

4. Related Work

Grouping PAM (Drinking Water Company) customers by province in Indonesia improves operational efficiency, service quality, and public service responsiveness. Through a deep understanding of customer characteristics in various regions, PAM can develop more effective business strategies and align its services with the needs of local communities. This research utilizes the K-Medoids algorithm to group PAM customers. This algorithm was chosen because of its superiority in dealing with outlier data and ease of interpretation of cluster results. Thus, using the K-Medoids algorithm is expected to provide a better understanding of water use distribution patterns in various provinces in Indonesia. PAM customer clustering has several significant benefits. First, this can improve the quality of services by providing solutions that are more targeted and tailored to the needs of local communities.

PAM can use information about customer characteristics to develop more effective strategies, such as improving infrastructure, expanding water networks, or improving water quality. Apart from that, customer grouping can also help PAM in developing more effective business strategies. By understanding the customer profile in each cluster, PAM can build products and services that better suit the needs of each market segment. This could include setting the correct prices, developing new products, or optimizing water distribution. Operational efficiency can also be improved through customer grouping. By identifying high-potential customers requiring special attention, PAM can allocate its resources more efficiently. For example, PAM can implement unique programs for loyal customers, provide subsidies for less fortunate customers, or launch educational programs for customers who waste water. Public service responsiveness can also be improved through customer grouping. By understanding customer complaints and needs in various regions, PAM can develop more effective strategies to deal with these problems. This could include improving the quality of call centers, expanding complaint channels, or training public service officers.

Several studies have been carried out to apply clustering algorithms in customer grouping. Simanjuntak *et al*. (2023) applied the K-Medoids algorithm for grouping unemployment in North Sumatra [1]. Anggarwati *et al*. (2021) use the K-Means algorithm to predict car body sales [2]. Pangestu and Ridwan (2022) apply the K-Means algorithm to group customers based on the volume of water sold [3]. Simanjuntak (2022) uses the clustering method to analyze customer satisfaction with PDAM services [4]. Fajriana (2021) analyzed the K-Medoids algorithm in the capture fisheries production clustering system [5]. Lestari *et al*. (2023) used the K-Means clustering algorithm for data mining on clean water sales at SPAM Akidah [6][5]. Ula *et al*. (2023) applied KNN to determine new PDAM customers and K-Means clustering based on region [7]. Triyandana *et al*. (2022) apply the K-Means method to group food and drink menus based on sales level [8]. Pahlawan *et al*. (2019) examined the influence of product and service quality on PAM customer satisfaction and loyalty [9]. Sinaga *et al*. (2019) applied data mining on the number of PAM customers by province using the K-Means clustering method [10]. Yaumi *et al*. (2020) clustered consumer characteristics toward product selection tendencies using K-Means [11]. Fajar Fauzan *et al*. (2023) applied data mining to analyze sales of bottled drinking water during the Covid-19 pandemic [12]. Ningsih *et al*. (2019) analyzed K-Medoids in grouping the illiterate population by province [13]. Meizar *et al*. (2022) analyzed trend moments in data mining forecasting in predicting the amount of herbal medicine inventory [14]. Siregar *et al*. (2020) carried out clustering on PAM customers using the K-Means algorithm [15]. Siska (2019) applies data mining with the Naive Bayes algorithm to customers for car service [16]. Andini and Arifin (2020) implemented the K-Medoids algorithm for clustering patient disease data at the Bandung City Regional Hospital [17]. Herdini and Widiyarta (2020) examined the responsiveness of public services in handling customer complaints at PAM Nganjuk Regency [18].

Kairupan *et al*. (2020) discuss PAM services in providing clean water. This research is expected to understand better the characteristics of PAM customers in various provinces in Indonesia and provide a solid

basis for making better water resource management decisions [19]. By using the K-Medoids algorithm and utilizing information obtained from customer groupings, PAM can improve operational efficiency, service quality, and responsiveness of their public services to be more effective in providing clean water for the people of Indonesia.

5. Conclusion

Based on the results of this research, the k-medoids algorithm can be applied to grouping PAM customers based on provinces in Indonesia. By comparing the results of manual calculations using Microsoft Excel on PAM customer data from 1998-2021, in which 34 provinces in Indonesia, 3 clusters were obtained with the same results. Cluster 1 (C1) is a low-level cluster that receives results from 19 provinces, each of which has results on the number of PAM customers, the amount of water distributed, and the value of the water distributed in low quantities. Cluster 2 (C2) is a medium-level cluster that obtains results from 11 provinces, each of which has results for the number of PAM customers, the amount of water distributed, and the value of the water distributed in moderate amounts. Cluster 3 (C3) is a high-level cluster that obtains results from 4 provinces, each with many PAM customers, the amount of water distributed, and the value of the water distributed. And get a k-medoids cluster validation value of 0.014846718. The cluster has good variance because it is close to 0; the smaller the cluster variance, the better. By applying the k-medoids algorithm in clustering PAM customers based on provinces in Indonesia, it is hoped that it can provide a better understanding of the clustering of PAM customers in each province and assist the government in making decisions regarding the management of clean water resources, as well as taking appropriate steps in improving customer service and satisfaction.

References

[1] Simanjuntak, D. S. M., Gunawan, I., Sumarno, S., Poningsih, P., & Sari, I. P. (2023). Penerapan Algoritma K-Medoids Untuk Pengelompokkan Pengangguran Umur 25 tahun Keatas Di Sumatera Utara. *J. Krisnadana, 2*(2). https://doi.org/10.58982/krisnadana.v2i2.264

[2] Anggarwati, D., Nurdiawan, O., Ali, I., & Kurnia, D. A. (2021). Penerapan Algoritma K-Means Dalam Prediksi Penjualan Karoseri. *J. Data Sci. Inform., 1*(2), 58–62. [Online]. Available: https://www.publikasi.bigdatascience.id/index.php/jdsi/article/view/32/15

[3] Pangestu, A., & Ridwan, T. (2022). Penerapan Data Mining Menggunakan Algoritma K-Means Pengelompokan Pelanggan Berdasarkan Kubikasi Air Terjual Menggunakan Weka. *JUST IT J. Sist. Informasi, Teknol. Inf. dan Komput., 11*(3), 67–71.

[4] Simanjuntak, A. D. (2022). Data Mining Kepuasan Pelanggan Terhadap Pelayanan Pdam Menggunakan Metode Clustering. *Semin. Nas. Inform., 6*(3). [Online]. Available: http://jurnal.kaputama.ac.id/index.php/SENATIKA/article/view/947/0%0Ahttp://jurnal.kaputama.ac.id/index.php/SENATIKA/article/viewFile/947/672

[5] Fajriana, F. (2021). Analisis Algoritma K-Medoids pada Sistem Klasterisasi Produksi Perikanan Tangkap Kabupaten Aceh Utara. *J. Edukasi dan Penelit. Inform., 7*(2), 263. https://doi.org/10.26418/jp.v7i2.47795

[6] Lestari, P. D., Kecomberan, D., Talun, K., Clustering, M., Clustering, A. K., & Pengguna, N. (2023). DATAMINING PADA PENJUALAN AIR BERSIH DI SPAM AKIDAH MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING, *7*(1), 412–416.

[7] Ula, M., Maulana, R., & Ilhadi, V. (2023). Terbit online pada laman web jurnal: https://ejurnalunsam.id/index.php/jicom/ Penerapan KNN Penentuan Pelanggan Baru PDAM dan Clustering K-Means Berdasarkan Wilayah, *04*(01), 1–6. [Online]. Available: https://ejurnalunsam.id/index.php/jicom/.

[8] Triyandana, G., Putri, L. A., & Umaidah, Y. (2022). Penerapan Data Mining Pengelompokan Menu Makanan dan Minuman Berdasarkan Tingkat Penjualan Menggunakan Metode K-Means. *J. Appl. Informatics Comput., 6*(1), 40–46. https://doi.org/10.30871/jaic.v6i1.3824

[9] Pahlawan, M. R., Nurlia, N., Laba, A. R., Pakki, E., & Hardiyono, H. (2019). Pengaruh Kualitas Produk Dan Kualitas Pelayanan Terhadap Peningkatan Kepuasan Dan Loyalitas Pelanggan Perusahaan Daerah Air Minum (Pdam) Kota Makassar. *J. Appl. Bus. Adm., 3*(2), 228–244. https://doi.org/10.30871/jaba.v3i2.1560

[10] Sinaga, L., Ahmad, A., & Safii, M. (2019). Penerapan Data Mining Pada Jumlah Pelanggan Perusahaan Air Bersih Menurut Provinsi Menggunakan Metode K-Means Clustering. *J. Resist. (Rekayasa Sist. Komputer), 2*(2), 119–125. https://doi.org/10.31598/jurnalresistor.v2i2.418.

[11] Yaumi, A. S., Zulfiqkar, Z., & Nugroho, A. (2020). Klasterisasi Karakter Konsumen Terhadap Kecenderungan Pemilihan Produk Menggunakan K-Means. *JOINTECS (Journal Inf. Technol. Comput. Sci., 5*(3), 195. https://doi.org/10.31328/jointecs.v5i3.1523.

[12] Fauzan, M. F., Purnamasari, A. I., & Dwilestari, G. (2023). Penerapan Data Mining Untuk Menganalisis Penjualan Air Minum Dalam Kemasan Selama Masa Pandemi Covid-19. *JATI (Jurnal Mhs. Tek. Inform., 7*(1), 700–706. https://doi.org/10.36040/jati.v7i1.6290.

[13] Ningsih, S. R., Damanik, I. S., Windarto, A. P., Tambunan, H. S., Jalaluddin, J., & Wanto, A. (2019). Analisis K-Medoids Dalam Pengelompokkan Penduduk Buta Huruf Menurut Provinsi. *Pros. Semin. Nas. Ris. Inf. Sci., 1*(September), 721. https://doi.org/10.30645/senaris.v1i0.78.

[14] Meizar, A., Fahrozi, W., Indra, E., & Saputra, M. (2022). Analisis Trend Moment Pada Datamining Forecasting Dalam Memprediksi Jumlah Persediaan Obat Herbal. *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA), 5*(2), 103–106. https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v5i2.2452.

[15] Siregar, A. R., Andani, S. R., & Saputra, W. (2020). Clustering Pada Pelanggan Perusahaan Air Bersih Dengan Algoritma K-Means. *BRAHMANA J. Penerapan Kecerdasan Buatan, 1*(1), 7–14. https://doi.org/10.30645/brahmana.v1i1.2.

[16] Siska, Y. (2019). Penerapan Data Mining Dengan Algoritma Naive Bayes Pelanggan Terhadap Pelayanan Servis Mobil (Studi Kasus : Katamso Service). *Maj. Ilm. INTI, 6*, 354.

[17] Andini, A. D., & Arifin, T. (2020). Implementasi Algoritma K-Medoids Untuk Klasterisasi Data Penyakit Pasien Di Rsud Kota Bandung. *J. Responsif Ris. Sains dan Inform., 2*(2), 128–138. https://doi.org/10.51977/jti.v2i2.247.

[18] Herdini, F., & Widiyarta, A. (2020). Responsivitas Pelayanan Publik Dalam Menangani Keluhan Pelanggan Di Perusahaan Daerah Air Minum (Pdam) Kabupaten Nganjuk. *10*(1), 54–75.

[19] Kairupan, S. B., Mokat, J. E., & Pakasi, K. M. (2020). Pelayanan Perusahaan Daerah Air Minum dalam Penyediaan Air Bersih di Kecamatan Pasan Kabupaten Minahasa Tenggara. *J. Kaji. Kebijak. dan Ilmu Adm. Negara (JURNAL Adm., 1*(2), 41–44. https://doi.org/10.36412/jan.v1i2.1636.