Implementation of Data Mining Using K-medoids Clustering Method for Determining Social Assistance Recipients

Ratnasari Nurzainun[1], Sinawati[2], and Denis Prayogi[3]*
[1,2] Information Systems, STMIK PPPKIA Tarakanita Rahmawati, Tarakan, Indonesia
[3] Informatics Engineering, STMIK PPPKIA Tarakanita Rahmawati, Tarakan, Indonesia
Email: 2050011@student.ppkia.ac.id[1], sinawati@ppkia.ac.id[2], denis@ppkia.ac.id[3]

Abstract

Regarding social assistance, the village head, "Lurah" of Selumit, is currently reviewing data on residents based on government-provided statistics on the capabilities of low-income families and the need for social assistance. Therefore, this study proposes K-medoid clustering to ensure that the assistance provided is appropriate. The study collected 62 data on social assistance recipients consisting of 6 criteria, namely employment, assets, income, jak (who are still dependents), home status, and home conditions. K-Medoids analysis using the Euclidean distance function with K=3 produces cluster 1 with 12 data, cluster 2 with 31 data, and cluster 3 with 19 data. The recipients prioritized to receive social assistance are the data in cluster 2 by calculating the average of the most considerable maximum weight value.

***Keywords:** Data Mining, K-Medoids Clustering, Euclidean Distance, social assistance.*

1. Introduction

In Indonesia, social assistance is important to improve people's welfare and combat poverty. Social assistance programs such as Bantuan Langsung Tunai (BLT) and Program Keluarga Harapan (PKH) aim to help those in need [1]. However, identifying accurate, efficient, and targeted assistance recipients often poses challenges. In this case, information technology can be a solution to increase the efficiency of welfare recipient selection. One method used in data mining to support this process is the clustering method, especially the K-Medoids algorithm.

K-medoids are a method for grouping data based on the proximity between data points. This method has been proven to be more stable than K-means, especially when dealing with data with outliers or uneven distribution [2][3]. Several previous studies have also shown the application of this method in various fields, including in determining recipients of social assistance and scholarship programs [4][5]. Therefore, applying the K-medoids algorithm in grouping social assistance recipients in Selumit Village, Tarakan City, North Kalimantan, can be a solution to increase accuracy and efficiency in determining recipients of assistance.

Although the application of data mining methods such as K-medoids has been widely used in various social programs [6][7], data grouping is The main challenge, which is not always easy and can be influenced by several socio-economic factors. In Selumit Village, there are obstacles to data access and adequate infrastructure, so determining social assistance recipients is often inaccurate and inefficient. Therefore, this study aims to apply the K-medoids method in grouping people in Selumit Village and determining more appropriate and objective social assistance recipients according to the established criteria so that it can contribute to improving the quality of services related to social assistance [8][9].

2. Research Methodology

2.1. System Analysis

The system analysis used to determine the criteria for social assistance recipients at Selumit Village, Tarakan City, uses the K-Medoids Clustering method for data grouping. Figure 1 shows several functions needed in the system to be created.

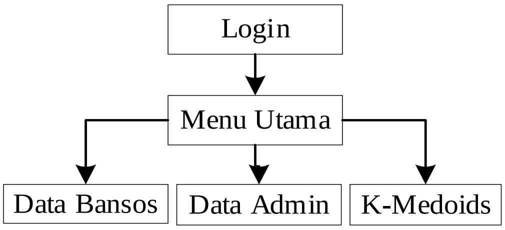


Figure 1. System Requirements Analysis

2.2. Data Collection

The data used for this study were 62 social assistance recipients taken at the Selumit Village Office, Tarakan City. The criteria used in this study were 6: employment, assets, income, dependents, house status, and house conditions (Hasanah), along with their assessment weights. Sample data can be seen in Table 1.

Table 1. Data for Determining Social Assistance Recipients

No.	**Employment (C1)**	**Assets (C2)**	**Income (C3)**	**Dependents (C4)**	**Home Status (C5)**	**Home Condition (C6)**
1	None	Motorcycle	500.000 to 1000.000	(1-3)	Freehold	Substandard
2	Laborer	No Assets	< 500,000	(4-6)	Occupied Temporarily	Uninhabitable
3	None	Motorcycle	No Income	(4-6)	Occupied Temporarily	Substandard
4	None	No Assets	No Income	(1-3)	Freehold	Uninhabitable
5	Laborer	Motorcycle	No Income	(1-3)	Freehold	Substandard
…	…	…	…	…	…	…
62	Laborer	No Assets	No Income	(1-3)	Rent	Habitable

Each criterion has a weight assessment from 1 to 6; the number represents the condition of each criterion. Table 2 explains the weight's value.

Table 2. assessment weight for each criterion

No.	**Criteria**	**Assessment Weight**				
		1	**2**	**3**	**4**	**5**
1	C1	Employee	Farmer	Fisherman	Laborer	None
2	C2	Land Certificate	Car	Motorcycle	Jewelry	No Assets
3	C3	>1.500.000	1.100.000 to 1.500.000	500.000 to 1000.000	< 500.000	No Income
4	C4	None	1-3	4-6	7-9	>9
5	C5	Freehold	Rental Rights (rented building)	Rent (boarding house)	Right to Use (only occupy the building)	Occupied Temporarily (stay at parents' or relatives' place)
6	C6	Very habitable	Habitable	Fairly habitable	Substandard	Uninhabitable

2.3. K-Medoids Algorithm

The stages in the K-Medoids algorithm are [10][11][12]:

a. Determine the desired k (number of clusters) from the data.
b. Randomly select the initial centers of k medoids from n data.
c. Calculate each object's distance to the temporary medoids using the Euclidean Distance calculation. The Euclidean Distance calculation formula is as follows.

$$d\ (x,y) = \sqrt{\sum_{i=1}^{n}(\boldsymbol{xi} - \boldsymbol{yi})^2} \qquad (1)$$

Description:
d(x,y) = Distance
xi = Coordinates of location 1
yi = Coordinates of location 2

d. Mark the closest distance from the object to the medoids and calculate the total.
e. Determine the cluster members to the temporary medoids.
f. Do iterations like Steps 2 to 4.
g. Calculate the total deviation (S)
   1. If a is the sum of the closest distances of the object from the initial medoids of iteration 1
   2. If b is the sum of the closest distances of the object from the initial medoids of iteration 2.
   3. The formula for the total deviation is:

$$s = b - a \qquad (2)$$

h. Total deviation (S)
   1. If S > 0 then the clustering process is stopped and members of each medoids are obtained.
   2. If S < 0 then exchange the object with other data to form a new set of k as medoids. Do iterations until the value of S> 0 is obtained.

3. Result and Discussion

Based on Tables 1 and 2, to perform calculations using K-Medoids, each data point on the criteria must be changed into a weighted number. Sample data can be seen in Table 3.

Table 3. Data for Determining Social Assistance Recipients

No.	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
Data1	5	3	3	2	1	4
Data2	4	5	4	3	5	5
Data3	5	3	5	3	5	4
Data4	5	5	5	2	1	5
Data5	4	3	5	2	1	4
…	…	…	…	…	…	…
Data62	4	5	5	2	2	2

Table 3 is a sample of data that has undergone a weighting process so that the criteria are represented numerically. From the k-medoid stage in point 2.3, the stages are:

a. The first step is to initialize the cluster center with as many as k numbers (clusters) and determine the initial medoid value selected randomly. It is assumed that the recipient data, namely Data4, Data15, and Data17, are medoids of iteration 1. So medoids 1 are (5,5,5,2,1,5), medoids 2 (5,5,5,2,5,2), and medoids 3 (4,3,4,2,1,5). The initial medoids data can be seen in Table 4.

Table 4. Initial Centroid Value

Medoids	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
Medoids 1 (Data4)	5	5	5	2	1	5
Medoids 2 (Data15)	5	5	5	2	5	2
Medoids 3 (Data17)	4	3	4	2	1	5

b. The second step calculates the distance between each object and the temporary medoids in Table 4.

1. Medoids 1 (Data4)

   This step calculates the distance between Data1 and Medoids 1, namely Data4.

   $$\sqrt{(5-5)^2+(3-5)^2\ +(3-5)^2+(2-2)^2+(1-1)^2+(4-5)^2}$$
   $$=3$$

2. Medoids 2 (Data15)

   This step calculates the distance between Data 1 and Medoids 2, namely Data15.

   $$\sqrt{(5-5)^2+(3-5)^2\ +(3-5)^2+(2-2)^2+(1-5)^2+(4-2)^2}$$
   $$=5{,}29$$

3. Medoids 3 (Data ke 17)

   This step calculates the distance between Data 1 and Medoids 2, namely Data17.

   $$\sqrt{(5-4)^2+(3-3)^2\ +(3-4)^2+(2-2)^2+(1-1)^2+(4-5)^2}$$
   $$=1{,}73$$

   Repeat this step by calculating all data using the initial Medoids so that the results can be seen in Table 5. From the calculation process of 62 data, the closest distance of each data is obtained. The closest distance value of each data is added up to get the total cost of iteration 1.

Table 5. Euclidean Distance Calculation at Iteration 1

No.	Euclidean Medoids 1	Euclidean Medoids 2	Euclidean Medoids 3	Shortest Distance	Cluster
Data1	3	5,29	1,73	1,73	3
Data2	4,36	3,46	4,58	3,46	2
Data3	4,69	3	4,47	3	2
Data4	0	5	2,45	0	1
Data5	2,45	5	1,41	1,41	3
…	…	…	…	…	…
Data62	3,32	3,16	3,74	3,16	2
Total Cost				116,31	

c. Initialize new cluster centers for the second iteration, which will be randomly selected. Assume the data selected for the new cluster centers are Data5, Data13, and Data32, as shown in Table 6.

Table 6. New Centroid Value

Medoids	(C1)	(C2)	(C3)	(C4)	(C5)	(C6)
Medoids 1 (Data5)	4	3	5	2	1	4
Medoids 2 (Data13)	5	5	5	3	5	4
Medoids 3 (Data32)	5	4	5	3	5	5

After determining the new centroid, the next step is to repeat the second step of calculating the distance and total cost.

d. The next step is to calculate the deviation value of the total cost from the current and previous iterations. The second terasi calculation using the Table 6 centroid produces a total cost value = 125.9. so that:

$$s\,=116.31\ -\ 125.9\ =\ -9.59$$

Because the value of S < 0, the clustering process is continued until the total deviation S > 0.

e. In this study, iterations occurred three times. The new centroids used in the third iteration were Data6 (5, 3, 5, 2, 5, 2), Data29 (5, 5, 5, 2, 1, 5), and Data41 (5, 5, 5, 3, 5, 4). By repeating the steps from number two to four, the last iteration produced a total cost value = 101.7. so that the value of S is:

$$s = 125.9 - 101.7 = 24.2$$

Because the value of S > 0, the clustering process is stopped. After that, the clustering process is carried out, and an evaluation is conducted to determine the priority cluster. Table 7 is a sample of clusters 1, 2, and 3 results.

Table 7. Cluster Results

No	C1	(C2)	(C3)	(C4)	(C5)	(C6)	Cluster
1	None	Motorcycle	No Income	(1-3)	Occupied Temporarily	Habitable	1
2	None	Motorcycle	No Income	(4-6)	Occupied Temporarily	Habitable	1
3	Laborer	Motorcycle	< 500,000	(1-3)	Occupied Temporarily	Very habitable	1
4	Laborer	No Assets	No Income	(1-3)	Freehold	Very habitable	1
5	None	No Assets	No Income	(1-3)	Freehold	Habitable	1
…	…	…	…	…	…	…	
13	None	Motorcycle	500.000 to 1000.000	(1-3)	Freehold	Substandard	2
14	None	No Assets	No Income	(1-3)	Freehold	Uninhabitable	2
15	Laborer	Motorcycle	No Income	(1-3)	Freehold	Substandard	2
16	None	No Assets	No Income	(4-6)	Freehold	Uninhabitable	2
17	None	No Assets	No Income	(1-3)	Freehold	Substandard	2
…	…	…	…	…	…	…	…
44	Laborer	No Assets	< 500,000	(4-6)	Occupied Temporarily	Uninhabitable	3
45	None	Motorcycle	No Income	(4-6)	Occupied Temporarily	Substandard	3
46	Laborer	Motorcycle	500.000 to 1000.000	(7-9)	Occupied Temporarily	Uninhabitable	3
47	Fisherma n	No Assets	500.000 to 1000.000	(7-9)	Occupied Temporarily	Uninhabitable	3
48	Laborer	No Assets	< 500,000	(4-6)	Occupied Temporarily	Uninhabitable	3
…	…	…	…	…	…	…	…

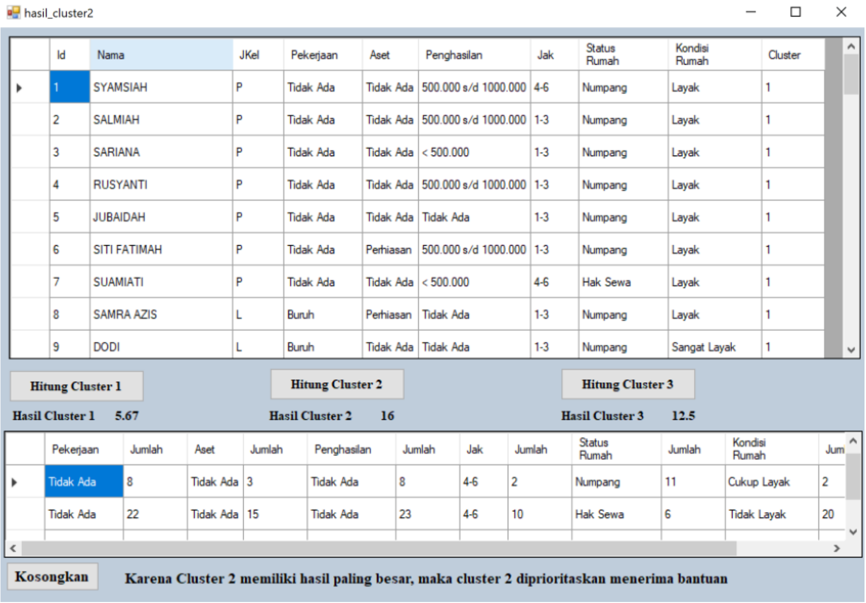
Table 7 shows a sample of results consisting of 3 clusters. Of the 62 recipient data, there are 12 data included in Cluster 1, 31 in Cluster 2, and 19 in Cluster 3. Calculate the sum of the maximum weight values of each criterion to determine which cluster will be a priority in receiving assistance. For example, in cluster 1 with the employment criterion, how many data do not have jobs because the weight is the highest? This aims to find the average with the highest value in the cluster to determine which one is prioritized. The results are shown in Table 8.

Table 8. Cluster Results for Determining Priorities

Maximum Criteria	Cluster 1	Cluster 2	Cluster 3
C1 (None)	8	22	15
C2 (No Assets)	3	15	13
C3 (No Income)	8	23	14
C4 (>9)	2	10	2
C5 (Occupied Temporarily)	11	6	19

Maximum Criteria	Cluster 1	Cluster 2	Cluster 3
C6 (Uninhabitable)	2	20	12
Average	**5.67**	**16**	**12.5**

Based on Table 8, the data prioritized for receiving assistance is the data in the second cluster because it has the most considerable maximum average value, namely 16. The appearance of the application created to assist the Sub-district in determining the provision of assistance can be seen in Figure 2.



hasil_cluster2

	Id	Nama	JKel	Pekerjaan	Aset	Penghasilan	Jak	Status Rumah	Kondisi Rumah	Cluster
▸	1	SYAMSIAH	P	Tidak Ada	Tidak Ada	500.000 s/d 1000.000	4-6	Numpang	Layak	1
	2	SALMIAH	P	Tidak Ada	Tidak Ada	500.000 s/d 1000.000	1-3	Numpang	Layak	1
	3	SARIANA	P	Tidak Ada	Tidak Ada	< 500.000	1-3	Numpang	Layak	1
	4	RUSYANTI	P	Tidak Ada	Tidak Ada	500.000 s/d 1000.000	1-3	Numpang	Layak	1
	5	JUBAIDAH	P	Tidak Ada	Tidak Ada	Tidak Ada	1-3	Numpang	Layak	1
	6	SITI FATIMAH	P	Tidak Ada	Perhiasan	500.000 s/d 1000.000	1-3	Numpang	Layak	1
	7	SUAMIATI	P	Tidak Ada	Tidak Ada	< 500.000	4-6	Hak Sewa	Layak	1
	8	SAMRA AZIS	L	Buruh	Perhiasan	Tidak Ada	1-3	Numpang	Layak	1
	9	DODI	L	Buruh	Tidak Ada	Tidak Ada	1-3	Numpang	Sangat Layak	1

Hasil Cluster 1 5.67

Hasil Cluster 2 16

Hasil Cluster 3 12.5

	Pekerjaan	Jumlah	Aset	Jumlah	Penghasilan	Jumlah	Jak	Jumlah	Status Rumah	Jumlah	Kondisi Rumah	Jum
▸	Tidak Ada	8	Tidak Ada	3	Tidak Ada	8	4-6	2	Numpang	11	Cukup Layak	2
	Tidak Ada	22	Tidak Ada	15	Tidak Ada	23	4-6	10	Hak Sewa	6	Tidak Layak	20

Karena Cluster 2 memiliki hasil paling besar, maka cluster 2 diprioritaskan menerima bantuan

Figure 2. Application View

4. Conclusion

Based on the description that has been explained, the implementation of K-Medoids to cluster prioritized aid recipient data was successfully carried out. The system built using Visual Basic programming shows that with K = 3, it produces cluster 2 with 31 data as a priority to receive social assistance from the government in Selumit Village, Tarakan City. In the future, development is needed in the form of more detailed and diverse criteria and further testing to evaluate the performance of K-Medoids in order to produce better clusters.

References

[1] R. Kinanti, Jasmir, and Fachruddin, "Penerapan Metode Clustering K-Means untuk Menentukan Prioritas Penerima Bantuan Program Beras Untuk Rakyat Miskin (Raskin) Studi Kasus : Kecamatan Siulak," *J. Inform. Dan Rekayasa Komput.*, vol. 4, no. 2, pp. 1135–1146, 2024, doi: 10.33998/jakakom.v4i2.

[2] I. Fatma, H. S. Tambunan, and F. Rizki, "Analisis Metode K-Medoids Cluster Dalam Mengelompokkan Siswa Yang Berprestasi," *Bull. Informatics Data Sci.*, vol. 1, no. 1, p. 14, 2022, doi: 10.61944/bids.v1i1.4.

[3] R. K. Purba and E. Bu'ulolo, "Implementasi Algoritma K-Medoids dalam Pengelompokan Mahasiswa yang Layak Mendapat Bantuan Uang Kuliah Tunggal (Studi Kasus : Universitas Budi Darma)," *INSOLOGI J. Sains dan Teknol.*, vol. 1, no. 2, pp. 79–86, 2022, doi: 10.55123/insologi.v1i2.195.

[4] M. A. Putri, A. Nazir, L. Handayani, and I. Afrianty, "Penerapan Algoritma K-Medoids Clustering untuk Mengetahui Pola Penerima Beasiswa Bank Indonesia Provinsi Riau," *JUKI J. Komput. ...*, vol. 5, no. 1, pp. 32–42, 2023, [Online].

Available: https://ioinformatic.org/index.php/JUKI/article/view/174%0Ahttps://ioinformatic.org/index.php/JUKI/article/download/174/155

[5] Y. F. Wijaya, “Implementasi Data Mining Untuk Penerima Bantuan PKH Pemerintah dengan Menerapkan Algoritma Klastering K-Medoids,” *J. Comput. Syst. Informatics*, vol. 5, no. 3, pp. 506–515, 2024, doi: 10.47065/josyc.v5i3.5197.

[6] R. N. H. Hutasuhut, H. Okprana, and B. E. Damanik, “Penerapan Data Mining untuk Menentukan Penerima Program Bidikmisi Menggunakan Algoritma K-Medoids,” *TIN Terap. Inform. Nusant.*, vol. 2, no. 11, pp. 667–672, 2022, doi: 10.47065/tin.v2i11.1516.

[7] Rispandi, “Implementasi Metode K-Medoids Untuk Clustring Penerima Bantuan Berdasarkan Normalisasi Data Masyarakat Miskin Dengan Metode Desimal Scaling,” *J. Comput. Informatics Res.*, vol. 3, no. 1, pp. 141–152, 2023, doi: 10.47065/comforch.v3i1.1063.

[8] S. Rahayu and A. Y. Kartini, “Algoritma K-Means dan K-Medoids untuk Pengelompokan Kecamatan Penerima Bantuan Sosial di Kabupaten Bojonegoro,” *Media Bina Ilm.*, vol. 16, no. 5, pp. 6815–6822, 2021.

[9] A. Iskandar, “Penerapan Algoritma K-Medoids Untuk Clustering Prioritas Penerima Beasiswa,” *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 508–514, 2023, doi: 10.47065/josh.v4i2.2927.

[10] N. T. Luchia, H. Handayani, F. S. Hamdi, D. Erlangga, and S. F. Octavia, “Perbandingan K-Means dan K-Medoids Pada Pengelompokan Data Miskin di Indonesia,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 35–41, 2022, doi: 10.57152/malcom.v2i2.422.

[11] H. Syukron, M. F. Fayyad, F. J. Fauzan, Y. Ikhsani, and U. R. Gurning, “Perbandingan K-Means K-Medoids dan Fuzzy C-Means untuk Pengelompokan Data Pelanggan dengan Model LRFM,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 2, no. 2, pp. 76–83, 2022, doi: 10.57152/malcom.v2i2.442.

[12] N. A. Kilo, M. R. Katili, and I. K. Hasan, “Perbandingan Metode K-Means dan K-Medoids Dengan Validitas Davies-Bouldin Indeks , Dunn Indeks dan Indeks Connectivity Pada Pengelompokkan Masyarakat Penerima Bantuan Langsung Tunai,” *Res. Math. Nat. Sci.*, vol. 4, no. 1, pp. 8–15, 2025, doi: 10.55657/rmns.v4i1.190.