Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2025. Vol. 4. No. 3; e-ISSN: 2808-4519

Grouping of Social Assistance Recipients Using K-Means Algorithm (Case Study: Gegunung Village Office)

Temu Asih[1]*, Rini Astuti [2], Willy Prihartono[3] , Ryan Hamonangan[4]

[1,3,4]Department of Informatics Engineering, STMIK IKMI Cirebon
[2] Department of Informatics Systems, STMIK LIKMI Bandung
temuasih270@gmail.com[1]*

Abstrak

The aim of this research is to use the K-Means algorithm to classify social assistance recipients in Gegunung Village based on location and nominal data. Inaccuracy and inefficiency are the main problems in the distribution of social assistance, so a technique is needed that can target grouping of recipient data. The Knowledge Discovery in Database (KDD) stage was used to process 672 data entries, including nominal information, location, occupation and type of assistance. Clusters were created using the K-Means method based on location and nominal value, and the Davies Bouldin Index (DBI) was used to assess quality. The findings show that six clusters with different data distributions were produced by optimal clustering with K=6 and DBI 0.971. Relevant parties can identify more effective distribution strategies with the help of these clusters, which provide insight into more structured social assistance distribution patterns. In short, the K-Means algorithm can be a useful tool for classifying social assistance data, facilitating more informed and effective decision making. This study significantly advances the domain of social assistance management and data collection.

***Keyword**: K-Means algorithm, Social assistance, Data mining, Clustering, Davies-Bouldin Index.*

1. Introduction

Social Assistance (BANSOS) is a government effort to channel assistance to individuals, families, or underprivileged communities to improve welfare, especially for those living below the poverty line. Programs such as the Family Hope Program (PKH) and Non-Cash Food Assistance (BPNT) require accurate data so that assistance is right on target. However, the distribution of social assistance still faces various obstacles, such as inaccurate data, uneven distribution, misappropriation of funds, and politicization of assistance.

One of the main challenges in social assistance programs is the effectiveness and accuracy in determining eligible recipients. Recipient data is often spread in large quantities with various economic and location attributes, but has not been optimally utilized. To overcome this problem, grouping social assistance recipients based on nominal and location using the K-Means algorithm can be a solution. This method is able to group people based on the type of assistance received, so that it can increase the accuracy and efficiency of social assistance distribution and ensure that assistance is received by people who really need it.

By implementing the K-Means algorithm, the government and social institutions can identify aid distribution patterns more systematically, allowing recipient segmentation based on economic factors and geographic location. This grouping can help in planning more targeted policies, reducing distribution inequality, and minimizing deviations in the distribution of social assistance. In addition, the use of data mining technology in managing social assistance can increase transparency and accountability, so that social assistance programs can run more efficiently and effectively in supporting community welfare.

2. Literatur Riview

2.1. Previous Studies

A study conducted by Sari Ufriani This study aims to facilitate the selection team to provide assistance according to the predetermined criteria whether or not they are eligible to receive the assistance. The data used in the study is 2020 data. Data processing in this study uses the K-Means Clustering method with 3 clusters, namely cluster 1 (C1) First priority, cluster 2 (C2) second priority and cluster 3 (C3) Third priority. In conducting the analysis, the author uses SPSS tools. The method used is the k-means clustering method with 1001 KK data, 5 attributes and 3 clusters. The number of clusters in manual calculations and using SPSS tools is, C1 contains 497 data, C2 contains 381 data, C3 contains 123 data [1].

Dwiguna & Bahtiar The main objective of this study is to build a Data Mining model using the K-means Clustering method to identify potential recipients of BLT assistance in Pamulihan Village. The results of the study obtained the best grouping is K9 with a Davies bouldin index (DBI) value of 0.745, this value is the most optimal because it is close to 0 and produces a data cluster with the largest number of data is cluster 0, which consists of 50 data consisting of 14 recipient data, 15 non-recipient data, and 21 data considered the least amount of data is cluster 2, which consists of 16 data, namely 15 recipients and 1 non-recipient [2].

Maulana Fathul Aziz This study aims to group PKH recipient data in Sidaharja Village by finding the best cluster. In addition, this study aims to obtain what Numerical Measures Type parameters produce the best group in PKH recipient data in Sidaharja Village. In this study, the K-means algorithm was used to obtain information from the results of grouping PKH recipient data in Sidaharja Village. The data analysis technique used Knowledge Discovery in Database (KDD). The results of the grouping show that from the number of clusters (K = 2, 3, 4, 5, 6) which produces the best K performance is the one that is close to 0. The K value closest to 0 is (K = 6). The optimal DBI is obtained for (K = 6) with a value of 0.230 for the Numerical Measures Type - Euclidean Distance parameter. To evaluate the performance of the K-means algorithm, it can be seen from the performance, where the Davies Bouldin value approaching 0 indicates the quality of the algorithm is getting better. Thus, the best number of clusters in this experiment is 6 or close to 0 [3].

Fitriyah The results obtained by modifying the Elbow method obtained the best clustering method is the K-Means method with 5 optimal clusters. The sub-district clusters in 2020 and 2021 are different because in 2020 there is more focus on COVID-19 pandemic assistance, while in 2021, there is more focus on direct Regional and Village assistance [4].

Edisman Rahul Gonjales Siahaan The aim of this is to apply the k-means method in grouping districts and cities on social welfare issues that can help the government and social services in making decisions which areas should be dominantly assisted in solving social welfare problems in order to save costs. The k-means method is one of the methods in data mining to group data sets that are similar to others. The data are grouped into 3 clusters, namely high, medium and low clusters, the results of the high cluster are 2 regencies/cities, the medium cluster is 6 regencies/cities and the low cluster is 25 regencies/cities, these results can be a record for the local government and agencies in dealing with social welfare problems in regencies and cities in North Sumatra [5].

1.2. Data Mining

Data mining is a set of procedures used to extract valuable insights, particularly previously unrecognized patterns, from datasets. This method identifies significant patterns in stored data within databases or retrieved datasets. Data mining is part of the Knowledge Discovery in Databases (KDD) process, aiming to extract meaningful information from large data warehouses [6].

1.3. Clustering

Clustering analysis is the process of dividing a dataset into groups, where data within the same group exhibit higher similarity compared to data in other groups. The potential of clustering lies in its ability to reveal data structures that can be applied to a variety of fields, such as classification, image processing, and pattern identification [7].

1.4. K-Means Algorithm

The K-Means algorithm is a data analysis technique used to group data based on patterns or behavioral characteristics. It begins by forming initial clusters and iteratively refines them until no significant changes occur. In this study, the K-Means algorithm was employed to classify patients based on variables such as age, type of substance abuse, and duration of use. Its application is expected to provide valuable insights for decision-making in drug rehabilitation programs [8].

1.5. RapidMiner

RapidMiner is a data science software platform developed by the company of the same name. It provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. The platform is used in business applications, research, education, training, prototyping, and application development. RapidMiner supports all stages of machine learning, including data preparation, result visualization, validation, and optimization, and is developed using an open-core model [9].

1.6. Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is a cluster validation method designed by D.L. Davies. DBI compares the ratio of within-cluster dispersion to between-cluster separation. Its goal is to maximize inter-cluster distance. In this study, DBI was applied as a validation metric to assess the quality of clustering. A clustering scheme is considered optimal if it achieves a Bouldin Index value [10].

3. Research Method

3.1. Data Source

Government agencies with a strong reputation for disseminating information on social assistance served as the data source for this study. The Gegunung Sub-district Office, the government agency tasked with overseeing and allocating social assistance in the area, provided most of the data. In addition, to refine and complete the primary data obtained from the Gegunung Village Office, this study utilized secondary data from a number of related supporting sources. For two weeks, data collection was carried out through direct observation. On October 14, 2024, an official meeting was held with the Gegunung Sub-district Head. During the meeting, the researcher obtained a dataset that included comprehensive data on 672 social assistance recipients. To facilitate analysis, this data was translated from its original xlsx file to csv format. The collected data contained a number of important details, including the recipients' identities, their financial situation, their fields of work, their income levels, and the types of assistance they received. initial screening procedures to verify the authenticity, consistency, and completeness of the data after it was collected.

3.2. Data Collection Techniques

The K-Means Clustering technique was used methodically to obtain data for this study. This method is used to ensure the quality, validity, and accuracy of the data to be examined. Direct observation and interviews were conducted at the research location to collect data including nominal information and the location of social assistance recipients in the Gegunung Village area.Data Analysis Techniques

1. Observation; To obtain the information needed in this study, the author conducted direct observation at the Gegunung Village Office from related sources.

2. Interview; In September and October 2024, the author conducted an interview stage by asking several questions related to Social Assistance to the Head of the Gegunung District Office or other Office officers.

3.3. Data Analysis Techniques

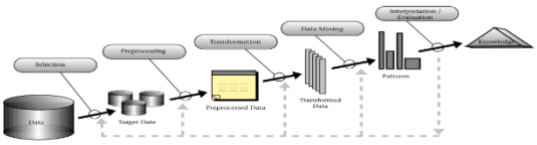


Fig. 1: Stages of the KDD Process

In this study, the K-Means algorithm was utilized for data clustering, following the stages of the Knowledge Discovery in Databases (KDD) process to uncover valuable patterns or information. The KDD process includes the following stages:

a. Data Selection; Data selection is the initial stage, where specific portions of data are chosen for further analysis. The success of the analysis largely depends on selecting the most relevant data.
b. Pre-Processing; Data cleaning is the process of detecting and correcting errors within the dataset. Its primary goal is to ensure high data quality so that the analysis can produce more accurate results.
c. Data Transformation; Data transformation involves converting data into a format that is more useful or suitable for further analysis. The aim is to improve data quality, enhance understanding, and facilitate the extraction of meaningful information.
d. Data Mining;Data mining is the process of discovering patterns or valuable insights from large datasets. At this stage, mathematical, statistical, and artificial intelligence techniques are employed to reveal hidden patterns.
e. Evaluation; Data interpretation is the process of assigning meaning or understanding to processed data. The goal is to identify significant patterns and use them for more accurate decision-making.

4. Discussion of Results

4.1. Analysis Results

The K-Means Clustering algorithm was implemented using the RapidMiner 10.0 tool:

4.1.1. Data Before Preprocessing

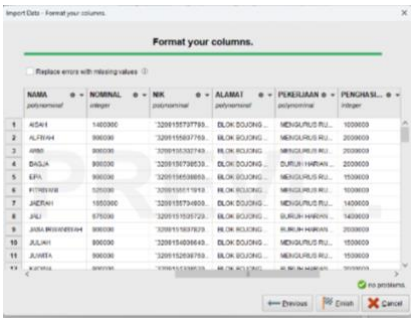


Fig. 2: Data Before Preprocessing

Figure 2 shows the pre-processed data obtained from the Gegunung Village social assistance data for October 2024.

4.1.2. Determining the Number of Clusters

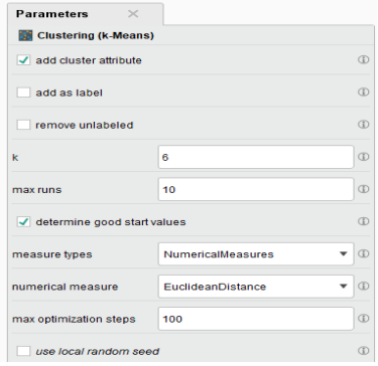


Fig. 3: Determining the Number of Clusters

Figure 3 illustrates the method for determining the value of k. The value of k is chosen based on the smallest Davies-Bouldin Index (DBI). In this data set, the smallest DBI value is achieved at k = 6 with a maximum run of 10. The comparison of DBI values is shown in table 1 below:

Table 1: Comparison of DBI Values

Value of K	DBI
2	1.258
3	1.256
4	1.069
5	0.977
6	0.971
7	1.020
8	1.031
9	1.031
10	1.029

Based on Table 1, the lowest value is observed in the cluster 6 with a DBI of 0.971.

4.1.3. Data Preprocessing

Fig. 4: Data Preprocessing

In Figure 4, the data displayed has undergone a process of selecting the attributes required for data processing in grouping Social Assistance based on certain characteristics.

4.1.4. Clustering Process in RapidMiner

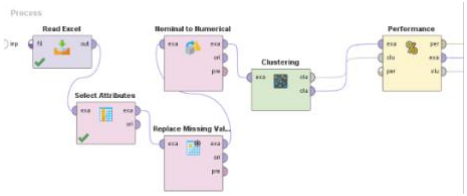
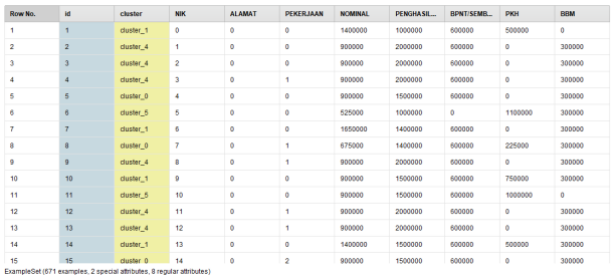


Fig. 5: Clustering Process in RapidMiner

Figure 5 depicts the data processing workflow using the K-Means Clustering algorithm in the RapidMiner application. This process utilizes six (6) operators:

1. Read Excel: Reads the medical record dataset from Puskesmas Jatiwangi.
2. Select Attributes: Selects the necessary attributes for the clustering process. The attributes used in this process include 2. *Select Attributes: Select the attributes required for the grouping process. The attributes used in this process include NIK, Address, Income, Nominal, Name, Type of assistance.*
3. Replace Missing Value: Handles missing data by removing duplicates and filling or managing missing values to ensure compliance with the required standards.
4. Nominal to Numerical: Converts non-numeric attributes into appropriate numerical representations.
5. K-Means Clustering: Performs clustering on the dataset with *K=6* and *Max Run=10*.
6. Cluster Distance Performance: Evaluates the performance of the K-Means Clustering algorithm on the processed dataset.

4.2. Clustering Results



Row No.	id	cluster	NIK	ALAMAT	PEKERJAAN	NOMINAL	PENGHASIL...	BPNT/SEMB...	PKH	BBM
1	1	cluster_1	0	0	0	1400000	1000000	600000	500000	0
2	2	cluster_4	1	0	0	900000	2000000	600000	0	300000
3	3	cluster_4	2	0	0	900000	2000000	600000	0	300000
4	4	cluster_4	3	0	1	900000	2000000	600000	0	300000
5	5	cluster_0	4	0	0	900000	1500000	600000	0	300000
6	6	cluster_5	5	0	0	525000	1000000	0	1100000	300000
7	7	cluster_1	6	0	0	1650000	1400000	600000	0	300000
8	8	cluster_0	7	0	1	675000	1400000	600000	225000	300000
9	9	cluster_4	8	0	1	900000	2000000	600000	0	300000
10	10	cluster_1	9	0	0	900000	1500000	600000	750000	300000
11	11	cluster_5	10	0	0	900000	1500000	600000	1000000	0
12	12	cluster_4	11	0	1	900000	2000000	600000	0	300000
13	13	cluster_4	12	0	1	900000	2000000	600000	0	300000
14	14	cluster_1	13	0	0	1400000	1500000	600000	500000	300000
15	15	cluster_0	14	0	2	900000	1500000	600000	0	300000

ExampleSet (671 examples, 2 special attributes, 8 regular attributes)

Fig. 6: Clustering Results Data in RapidMiner

Figure 6 presents the clustering result data processed in RapidMiner.

4.2.1. Cluster Model

Figure 7 illustrates the results of this study, which produced six (6) clusters with varying numbers of items. Cluster 0 contains 152 items, Cluster 1 contains 80 items, Cluster 2 33 items, Cluster 3 82 items, Cluster 4 219 items, while Cluster 5 105 items. The total number of items in the 6 clusters is 671.

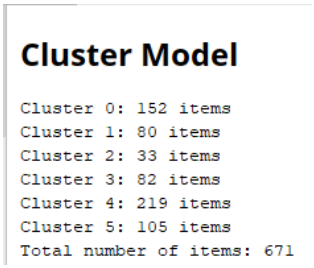


Fig. 7: Cluster Model

4.2.2. Social Assistance Cluster Results

Berikut ini adalah hasil pengelompokan dari data Bantuan Sosial seperti pada Gambar 8 yang menampilkan hasil pengelompokan dari penerapan algoritma K-Means dengan menggunakan tipe plot Scatter/Bubble.

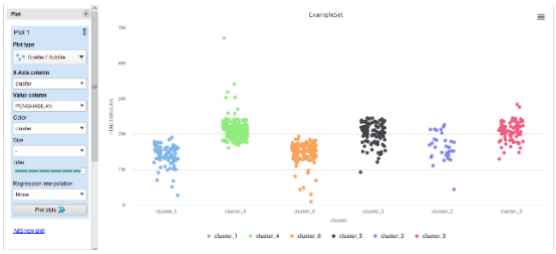


Fig. 8: Cluster Results of Disease Types in Medical Records

1. **Cluster 0**, contains 27 blocks or addresses, each of which represents a number of related data or information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by income factors. In addition, in this cluster there are 152 items representing various income values with the analysis being carried out.
2. **Cluster 1**, consists of 21 blocks or addresses, each of which represents a certain amount of data or related information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by the income factor. In addition, in this cluster there are 80 items that represent various income values with the analysis being carried out.

3. **Cluster 2,** consists of 14 blocks or addresses, each of which represents a certain amount of data or related information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by the income factor. In addition, in this cluster there are 33 items that represent various income values with the analysis being carried out.
4. **Cluster 3,** consists of 16 blocks or addresses, each of which represents a certain amount of data or related information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by the income factor. In addition, in this cluster there are 82 items representing various income values with the analysis being carried out.
5. **Cluster 4,** consists of 32 blocks or addresses, each of which represents a certain amount of data or related information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by income factors. In addition, in this cluster there are 219 items representing various income values with the analysis being carried out.
6. **Cluster 5,** consists of 24 blocks or addresses, each of which represents a certain amount of data or related information. The number of blocks is obtained based on the nominal value calculated and is greatly influenced by the income factor. In addition, in this cluster there are 105 items representing various income values with the analysis being carried out.

5. Conclusion

The research conducted in Gegunung Village, using 671 Social Assistance data, showed that the optimal K value for the K-means algorithm was 6, with a DBI of 0.971, indicating a good clustering evaluation. Further research is recommended to use larger and more diverse datasets to improve accuracy, and to explore other clustering algorithms to compare performance and determine the best method. Combining a wider dataset with different algorithms can provide deeper insights into Social Assistance, thereby increasing the reliability of research findings.

Acknowledgement

All praise and gratitude are due to Allah SWT for His abundant grace, blessings, and guidance, which enabled the author to complete this research successfully. This research would not have been possible without the support, assistance, and guidance of many parties. Therefore, with the utmost respect and sincerity, the author extends heartfelt thanks to:

1. Assoc. Prof. Dr. Dadang Sudrajat, S.Si., M.Kom, as the Chairman of STMIK IKMI Cirebon.
2. Mr. Dian Ade Kurnia, M.Kom, as Vice Chairman I for Academic Affairs, Collaboration, Research, and Innovation.
3. Mrs. Dra. Nining R, M.Si., as Vice Chairman II for Finance.
4. Mrs. Fatihanursari Dikananda, S.Tr.I.Kom., M.Kom, as Vice Chairman III for Student Affairs and Alumni.
5. Mr. H. Eka Jayawangsa, BBA, as Vice Chairman IV for Facilities and Infrastructure.
6. Mrs. Gifthera Dwilestari, S.I.Kom., M.Kom, as the Head of the Informatics Engineering Study Program.
7. Mrs. Rini Astuti, MT, as the Principal Supervisor.
8. Mr. Willy Prihartono, M.Kom, as the Co-Supervisor.
9. My beloved parents and family, who have consistently provided unwavering support, prayers, and encouragement throughout this academic journey.
10. My friends and all parties who have contributed, directly or indirectly, to the completion of this research.

May all the kindness and support extended be rewarded manifold by Allah SWT.

References

[1] Sari Ufriani, Jasmir, and Y. Arvita, "Penerapan Algoritma Clustering K-means untuk Menentukan Prioritas Penerima Bantuan Dana Sosial PKH di Kelurahan Kampung Singkep," *J. Inform. Dan Rekayasa Komputer(JAKAKOM)*, vol. 3, no. 1, pp. 342–350, 2023, doi: 10.33998/jakakom.2023.3.1.726.

[2] E. Dwiguna and A. Bahtiar, "Penerapan Data Mining Untuk Menentukan Penerima Bantuan Blt Menggunakan Metode Clustering K-Means Pada Desa Pamulihan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1382–1388, 2024, doi: 10.36040/jati.v8i2.9029.

[3] D. Maulana Fathul Aziz, B. Irawan, and A. Bahtiar, "Analisis Dataset Program Keluarga Harapan (Pkh) Desa Sidaharja Menggunakan Algoritma K-Means," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3370–3376, 2024, doi: 10.36040/jati.v7i6.8200.

[4] H. Fitriyah, E. M. Safitri, N. Muna, M. Khasanah, D. A. Aprilia, and D. Nurdiansyah, "Implementasi Algoritma Clustering Dengan Modifikasi Metode Elbow Untuk Mendukung Strategi Pemerataan Bantuan Sosial Di Kabupaten Bojonegoro," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 4, no. 3, pp. 1598–1607, 2023, doi: 10.46306/lb.v4i3.453.

[5] Edisman Rahul Gonjales Siahaan, "Implementation Of The K-Means Method In Grouping Districts And Cities In North Sumatra On Social Welfare Problems," *J. Artif. Intell. Eng. Appl.*, vol. 1, no. 2, pp. 168–173, 2022, doi: 10.59934/jaiea.v1i2.86.

[6] F. Fitriani, R. Kurniawan, and T. Suprapti, "Penerapan Algoritma K-Means Clustering Untuk Identifikasi Kelayakan Penerima Bantuan Program Keluarga Harapan (Pkh) Di Desa Tambaksari Ciamis," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 6, pp. 3363–3369, 2024, doi: 10.36040/jati.v7i6.8197.

[7] D. Setiawan, D. Arsa, L. E. Fitri, F. Fadhila, and P. Zahardy, "Comparative Analysis of Clustering Approaches in Assessing ChatGPT User Behavior," vol. 10, no. 2, pp. 366–379, 2024.

[8] W. R. Rahayu, A. I. Purnamasari, and A. Bahtiar, "Improving Regional Clustering Based on Tuberculosis Cases using the K-Means Algorithm of the Cirebon City Health Office," vol. 4, no. 2, 2025.

[9] I. Dinda Anjani and A. Bahtiar, "Penerapan Algoritma K-Means Clustering Untuk Mengelompokkan Penerima Bantuan Sosial Tunai (Bst) Di Jawa Barat," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 2743–2747, 2024, doi: 10.36040/jati.v8i3.8974.

[10] A. Adiyanto and Y. Arie Wijaya, "Penerapan Algoritma K-Means Pada Pengelompokan Data Set Bahan Pangan Indonesia Tahun 2022-2023," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1344–1350, 2023, doi: 10.36040/jati.v7i2.6849.