Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 12 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i12.42147

PAPER

Optimizing Patient Medical Records Grouping through Data Mining and K-Means Clustering Algorithm: A Case Study at RSUD Mohammad Natsir Solok

Dony Novaliendry[1](✉), Tegar Wibowo[1], Noper Ardi[2], Tiolina Evi[3], Dwi Admojo[3]

[1]Universitas Negeri Padang, Padang, Indonesia

[2]Politeknik Negeri Batam, Batam, Indonesia

[3]Perbanas Institute, Jakarta, Indonesia

dony.novaliendry@ft.unp.ac.id

ABSTRACT

RSUD Mohammad Natsir Solok, located in Solok City, provides comprehensive individual health services within its premises, offering both inpatient and outpatient care with 24-hour service availability. Inpatient services encompass emergency care and basic health services. A crucial component of healthcare operations is medical records, which consist of documented information pertaining to patient identity, examinations, treatments, procedures, and other services rendered. Medical records are essential and should be meticulously created in written or electronic form to ensure completeness and clarity. One common challenge encountered in maintaining medical records is the presence of overlapping data. To tackle this issue, data mining techniques are employed, with clustering being the primary method of choice. The K-Means algorithm is specifically utilized for clusterization purposes. By applying this data mining process and grouping patient medical records, valuable insights into the patterns of disease spread across different villages can be obtained. After applying K-Means clustering method, four distinct clusters were identified. The first cluster comprises 562 items, the second has 406 items, and the third and fourth have 791 and 279 items, respectively. These findings can serve as a reference for the local government, particularly the Solok City Health Office, to facilitate disease prevention initiatives and awareness campaigns. Decision-making related to disease sources, diagnosis, age, and gender of the affected patient can be informed by this data analysis.

KEYWORDS

medical records, clustering, K-Means

1 INTRODUCTION

RSUD Mohammad Natsir is a reputable hospital located in Solok City, West Sumatra Province [1]. Specifically situated in the Simpang Rumbio sub-district,

Novaliendry, D., Wibowo, T., Ardi, N., Evi, T., Admojo, D. (2023). Optimizing Patient Medical Records Grouping through Data Mining and K-Means Clustering Algorithm: A Case Study at RSUD Mohammad Natsir Solok. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(12), pp. 144–155. https://doi.org/10.3991/ijoe.v19i12.42147

Article submitted 2023-05-03. Resubmitted 2023-06-08. Final acceptance 2023-06-10. Final version published as submitted by the authors.

© 2023 by the authors of this article. Published under CC-BY.

Lubuk Sikarah District, Solok City [2], the hospital serves as a vital healthcare facility in the area. The hospital offers comprehensive healthcare services, including inpatient and outpatient services, with 24-hour availability. The inpatient services provided by RSUD Mohammad Natsir encompass emergency care and basic healthcare, ensuring that patients receive immediate and appropriate medical attention within the hospital premises. Additionally, the hospital is actively involved in public health initiatives, extending its services beyond the confines of the hospital building. The public health services offered by RSUD Mohammad Natsir encompass a wide range of programs aimed at promoting community well-being. These programs include health promotion activities, disease eradication efforts, environmental health initiatives, nutrition improvement campaigns, family health improvement projects, family planning services, mental health support, and various other public health programs.

Medical records play a vital role in documenting and storing all relevant information and documents related to a patient's health condition [3] [4]. These records consist of essential details such as patient identity, examination records, treatment information, medical interventions, and other services provided. It is crucial for medical records to be complete, clear, and accurately documented, whether in written or electronic form. The administration of medical records, particularly when utilizing electronic information technology, is governed by specific regulations. These regulations ensure the proper handling and management of sensitive patient information. Doctors, health workers, management officers, and healthcare facility heads are obligated to treat the information contained in medical records as confidential. As hospital activities progress, medical records data continues to accumulate daily. Medical records serve various purposes, including (1) facilitating health maintenance and treatment of patient, (2) providing evidence in legal proceedings, medical discipline, and ethical enforcement in medicine and dentistry, (3) fulfilling educational and study requirements, (4) serving as a basis for health services payments, (5) generating health statistical data. One common issue encountered in medical records management is the problem of data overlap among different records within hospitals [5].

One way to address this challenge is through the utilization of data mining. The field of Data Mining Science has introduced innovative approaches in order to leverage large datasets, enabling the extraction of valuable knowledge, both within specific domains and on a global scale. Data mining science encompasses various functions, including estimation, prediction, clustering, classification, and association. These functions are achieved through the utilization of different methods or algorithm, such as regression for estimation, Support Vector Machine (SVM) for prediction, K-Means for clustering, C4.5 for classification, and a priori for association [6] [7].

K-Means *Clustering* method operates on the principle of the grouping objects based on proximity measures and using the subsequent characteristics as centroids or feature vectors. This technique offers a viable solution for classifying the characteristics of objects and demonstrates a relatively high level of accuracy, particularly in object sizing. With this in mind, the current study aims to explore the patterns or rules that can be used for classifying patient medical records data using K-Means algorithm.

2 THEORETICAL BASIS

2.1 Data mining

Data mining is the process of extracting valuable information from large databases to facilitate decision-making [8]. Furthermore, it is a scientific discipline that aims to discover and explore knowledge from available data or information. The characteristics of data mining can be summarized as follows [9].

1. It involves uncovering hidden data patterns that were previously unknown.
2. It typically utilizes extensive datasets, often referred to as big data, to enhance the credibility of the results.
3. It plays a crucial role in making critical decisions, especially in a strategic context.

2.2 K-Means clustering

K-Means Machine Learning algorithm is the most popular and widely used clustering algorithm. It operates iteratively, aiming to partition dataset into pre-defined clusters, with each data point assigned to a single group [9]. Among the techniques employed in data mining, clustering is a prominent one. Scientific clustering in data mining involves grouping a set of data or objects into clusters, where each cluster contains similar data while being distinct from others [10]. The algorithm has gained significant attraction across various industries. It starts by selecting K initial cluster centers from the dataset and proceeds to assign each sample to the nearest cluster center based on their proximity [11][12]. It is important to acknowledge that this algorithm is a non-hierarchical method that partitions objects into one or more clusters based on their characteristics. Objects with similar characteristics are grouped, while those with different characteristics are placed in separate clusters [13][14].

2.3 RapidMiner

RapidMiner is a software application used as a learning tool in data mining science. The platform was developed by a company dedicated to handling large amounts of data in various fields such as business, commerce, education, training, and learning. It offers approximately 100 learning solutions for grouping, classification, and regression analysis [15]. Furthermore, the software is built for analytics teams and it integrates the entire lifecycle of data science, from data preparation to learning engines to model deployment. Over 625,000 analytical professionals use the products of the application to increase revenue, reduce costs, and avoid risks [16]. RapidMiner provides a UI for designing analysis pipelines and generates an XML file describing the analysis process that the user wants to apply to data. The software can read this file to run the analysis automatically [17].

3 STUDY METHOD

This study was conducted following a systematic flow including literature studies, data collection, data pre-processing, clustering with K-Means Clustering, and conclusions. The main objective was to cluster inpatient diagnostic data using the following stages (1) data analysis process, (2) K-Means Clustering calculations, and (3) testing with RapidMiner data mining tools.

3.1 Data analysis process

During the data analysis process, inpatient diagnosis data was used from the Mohammad Natsir Solok Hospital. The data is clustered using the K-Means Clustering algorithm, leading to model production. Furthermore, the model was analyzed using test data [18] as shown in Figure 1.

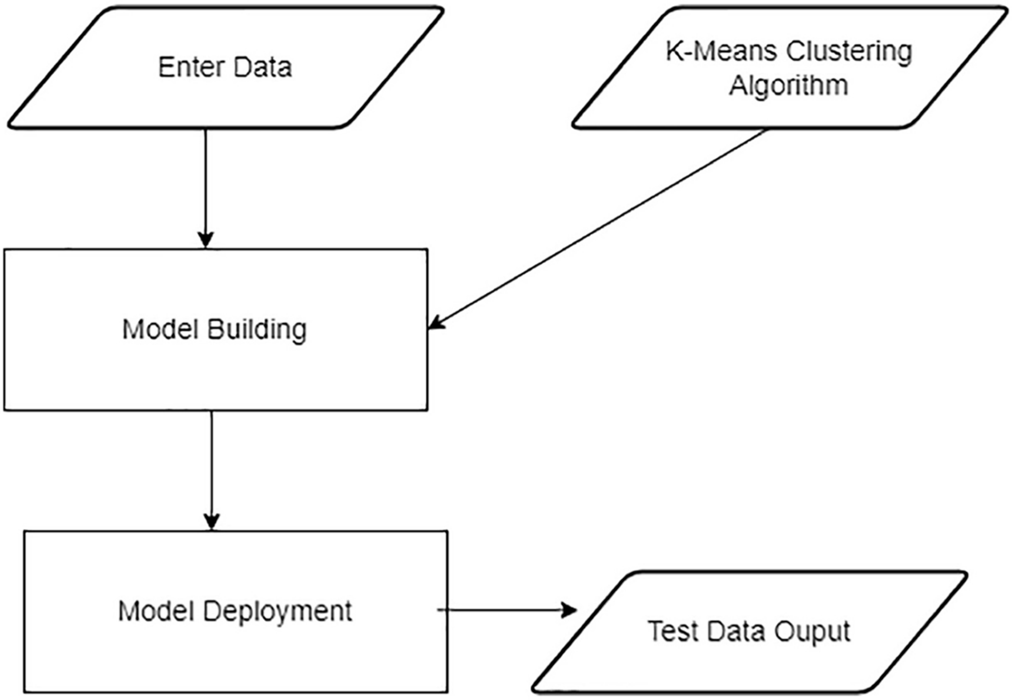


Fig. 1. Data analysis process

3.2 Calculation of K-Means clustering

The K-Means algorithm performs the following steps iteratively until stability is achieved or when there are no objects that can be moved [19], as shown in Figure 2 below.

1. Determine the coordinates of the midpoint for each cluster.
2. Determine the distance between each object and the midpoint coordinates.
3. Group the objects based on their minimum distance.

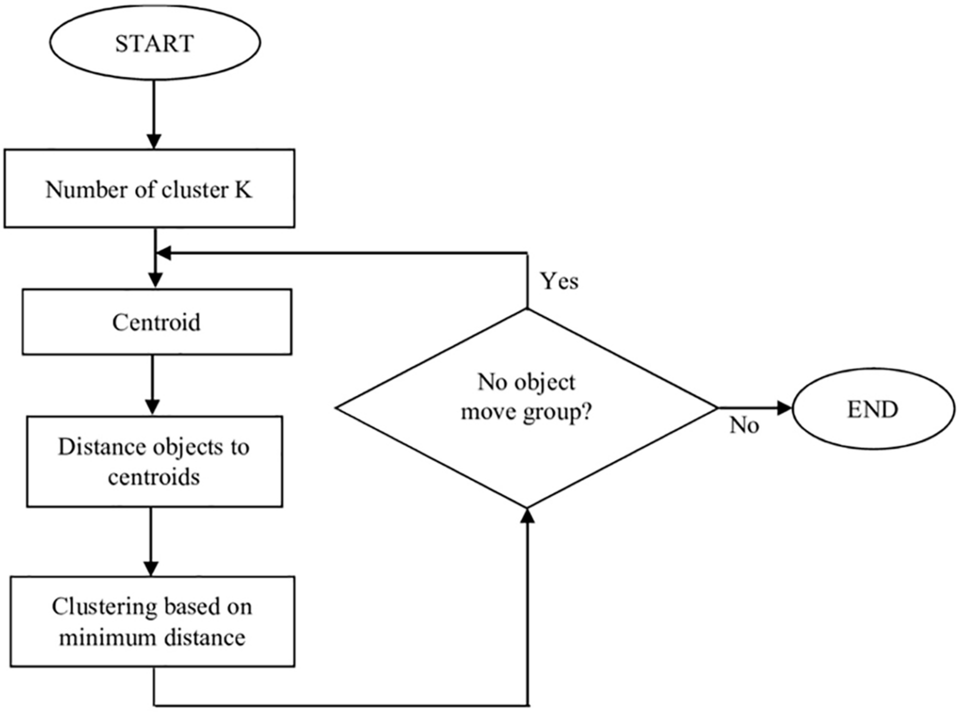

START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

Fig. 2. Flowchart K-Means algorithm

3.3 Calculating data distance with the euclidean distance formula

The stages involved in K-Means clustering calculation included (a) determining the number of clusters, (b) allocating data according to the number of clusters, (c) determining the centroid value for each cluster (d) calculating the shortest distance using the Euclidean distance formula.

$$D_{(I,j)} = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + \ldots + (Xki - Xkj)^2} \tag{1}$$

where :
$D_{(I,j)}$ = Data distance to i central cluster j
Xki = Data I on attribute data to k
Xkj = Center point jo j on attribute k

(e) Displaying results based on the lowest distance obtained from the calculation results of step 4. (f) If the same results had not been obtained, the iteration was repeated using step 3. (g) The iteration was stopped when clustering results were the same as the previous iteration.

3.4 Calculating new clusters

The calculation of the new cluster center was performed by calculating the average value of each criterion for all members belonging to each cluster.

$$\mu j(t+1) = \frac{1}{Nsj} \sum_{j \ni sj} xj \tag{2}$$

where:
μj (t + 1): newcenteroidat-(t+1)
Nsj: lots of data on cluster Sj

4 RESULT AND DISCUSSION

4.1 Clustering results

The iteration process ended in the 4th stage and produced 562 data in the first cluster, 406 in the second, 791 in the third, and 279 in the fourth. A sample of medical records data used could be seen in Tables 1–4.

Table 1. First cluster results

No.	Data to-	Cluster
1	1	Cluster 1
2	5	Cluster 1
3	9	Cluster 1
4	15	Cluster 1
5	21	Cluster 1
6	25	Cluster 1
7	31	Cluster 1
8	35	Cluster 1
9	37	Cluster 1
10	38	Cluster 1
:	:	:
562	2036	Cluster 1

Table 1 showed the cluster results obtained after an iteration, containing 562 data in the first cluster.

Table 2. Second cluster results

No.	Data to-	Cluster
1	2	Cluster 2
2	4	Cluster 2
3	7	Cluster 2
4	12	Cluster 2
5	14	Cluster 2
6	16	Cluster 2
7	17	Cluster 2
8	20	Cluster 2
9	26	Cluster 2
10	27	Cluster 2
:	:	:
406	2038	Cluster 2

Table 2 showed the cluster results obtained after iteration, with about 406 data in the second cluster.

Table 3. Third cluster results

No.	Data to-	Cluster
1	3	Cluster 3
2	6	Cluster 3
3	8	Cluster 3
4	13	Cluster 3
5	19	Cluster 3
6	22	Cluster 3
7	23	Cluster 3
8	24	Cluster 3
9	28	Cluster 3
10	34	Cluster 3
:	:	:
791	2037	Cluster 3

Table 3 showed the cluster results obtained after an iteration, containing as many as 791 data in the third cluster.

Table 4. Fourth cluster results

No.	Data to-	Cluster
1	10	Cluster 4
2	12	Cluster 4
3	18	Cluster 4
4	44	Cluster 4
5	47	Cluster 4
6	56	Cluster 4
7	69	Cluster 4
8	74	Cluster 4
9	77	Cluster 4
10	95	Cluster 4
:	:	:
279	2026	Cluster 4

Table 4 showed the cluster results obtained after iteration, containing approximately 279 data in the fourth cluster.

4.2 Testing with RapidMiner tools

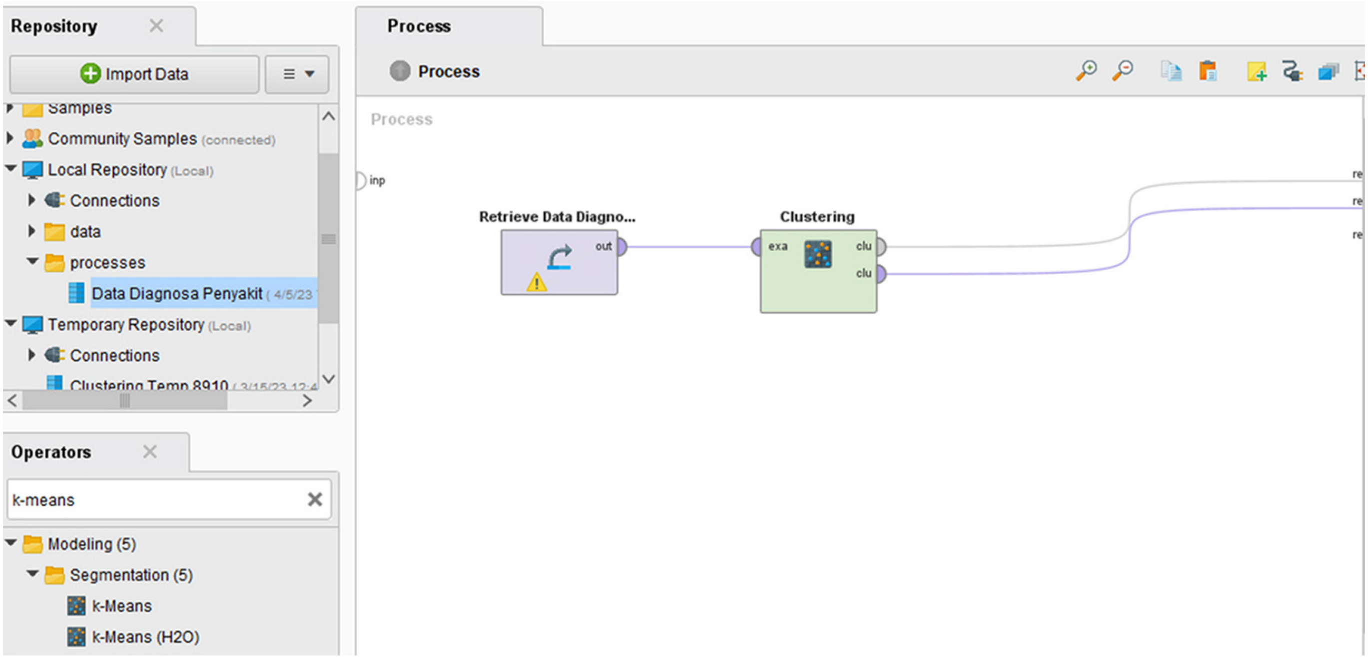
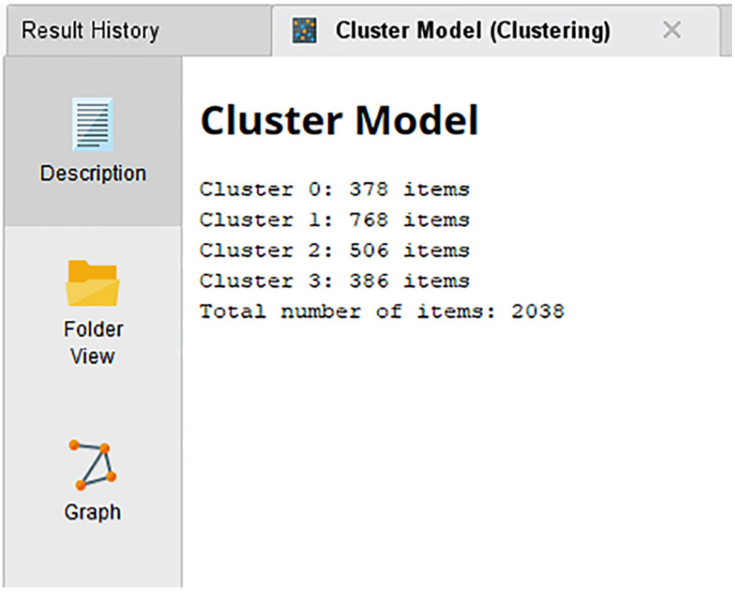


Fig. 3. Data processing process

The image above (Figure 3) shows data processing using the K-Means clustering algorithm, which produces the following output data.



Cluster Model

```
Cluster 0: 378 items
Cluster 1: 768 items
Cluster 2: 506 items
Cluster 3: 386 items
Total number of items: 2038
```

Fig. 4. Cluster model results

From the picture above (Figure 4), it could be observed that the cluster model grouped a total of 2038 data into 4 clusters. Clusters 1, 2, 3, and 4 consisted of 378, 768, 506, and 386 data, respectively.

4.3 Evaluating clusters

To determine the optimal number of clusters, the Davies Bouldin Index (DBI) technique was employed. The DBI technique, introduced by David L. Davies and Donald W. Bouldin in 1979, served as a metric for evaluating the results of clustering algorithm [20]. Based on this criterion, clustering that yielded a cluster set with the lowest Davies-Bouldin index was considered the best algorithm [21]. An experimental calculation of 4 to 7 clusters was performed using the RapidMiner tool, as shown in Figure 5.

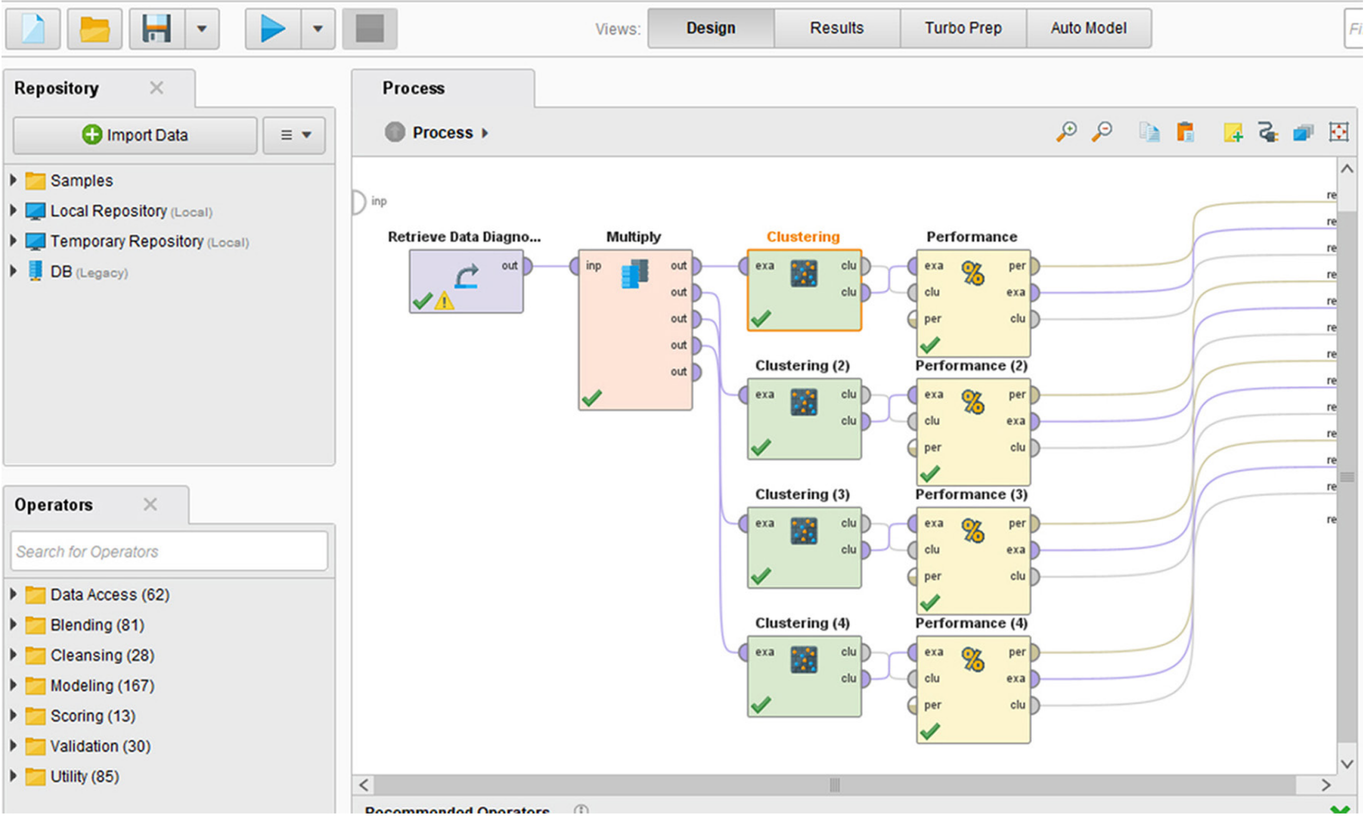


Fig. 5. The process of calculating the DBI value

After being processed in the RapidMiner tool, the DBI value was presented in Table 5.

Table 5. DBI value

Cluster Set	DBI Value
4	0.441
5	0.525
6	0.550
7	0.445

From the calculation results in the table, the cluster set with the lowest DBI value of 0.441 was the one with 4 clusters. The characteristics of each cluster were as follows:

The first cluster (C1) was characterized by a diagnosis of abdominal colic with the ICD-10 r104 code. It primarily affected adults, particularly females, in the Pasar Pandan Airmati sub-district.

The second cluster (C2) was characterized by a diagnosis of ischemic stroke with the ICD-10 code i639. It predominantly affected elderly males in the Pasar Pandan Airmati village.

The third cluster (C3) was characterized by a diagnosis of Dengue Haemorrhagic Fever (DHF) with the ICD-10 code a91. It mostly affected adults, particularly females, in the Tanah Garam sub-district.

The fourth cluster (C4) was characterized by a diagnosis of Chronic Kidney Disease (CKD) with the ICD-10 code n185. It mainly affected adults, particularly females, in the Simpang Rumbio sub-district.

5 CONCLUSION

Based on data analysis and testing, the following conclusions were drawn:

1. The determination of the centroid (central point) in the early stages of K-Means algorithm significantly influenced the cluster results. Tests were conducted using 2038 datasets with different centroids, producing different cluster results.
2. By employing K-Means clustering method and RapidMiner tools to classify patient medical records data, four clusters were formed. These methods identified diseases and sub-districts that often appeared, such as Dengue Hemorrhagic Fever (DHF), also known as dengue fever and PPA Village.
3. The grouping of patient medical records data through the aforementioned data mining process aimed to generate new insights into the pattern of disease distribution in each village. This information could serve as a reference for the local government, particularly Solok City Health Office, to conduct awareness campaigns and implement preventive measures targeting disease sources, considering the prevalent diseases, age groups, and gender of the affected patient.

6 REFERENCES

[1] Arwin, P., Arta, A., & Adhyka, N. (2022). Pelaksanaan Pos Gizi pada Masa Pandemi COVID-19 di Posyandu Kasih Ibu, RSUD M. Natsir Solok, Tahun 2020. Jurnal Inovasi, Pemberdayaan Dan Pengabdian Masyarakat, 2(1), 13–17. https://doi.org/10.36990/jippm.v2i1.466

[2] Kinerja, L. (2020). RSUD Mohammad Natsir.

[3] Hernawan, H., & Ningsih, K. P. (2020). Analisis Desain Map Rekam Medis. Jurnal Rekam Medis Dan Informasi Kesehatan, 3(2), 99–105. https://doi.org/10.31983/jrmik.v3i2.6331

[4] Wandana, J., Defit, S., & Sumijan, S. (2020). Klasterisasi Data Rekam Medis Pasien Pengguna Layanan BPJS Kesehatan Menggunakan Metode K-Means. Jurnal Informasi Dan Teknologi, 2(4), 4–9. https://doi.org/10.37034/jidt.v2i4.73

[5] Ordila, R., Wahyuni, R., Irawan, Y., & Yulia Sari, M. (2020). Penerapan Data Mining Untuk Pengelompokan Data Rekam Medis Pasien Berdasarkan Jenis Penyakit Dengan Algoritma Clustering (Studi Kasus : Poli Klinik PT.Inecda). Jurnal Ilmu Komputer, 9(2), 148–153. https://doi.org/10.33060/JIK/2020/Vol9.Iss2.181

[6] Ge, Z., Song, Z., Ding, S.X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access, 5, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

[7] Suyanto. (2017). Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informaatika.

[8] Novaliendry, D., Hendriyani, Y., Yang, C. H., & Hamimi, H. (2015). The Optimized K-Means Clustering Algorithms to Analyzed the Budget Revenue Expenditure in Padang. International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2, 61–66. https://doi.org/10.11591/eecsi.v2i1.771

[9] Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru. Sistemasi, 7(3), 238. https://doi.org/10.32520/stmsi.v7i3.388

[10] Guojun Shi, Bingkun Gao, & Li Zhang. (2013). The Optimized K-Means Algorithms for Improving Randomly-Initialed Midpoints. 2nd International Conference on Measurement, Information and Control. 2013:1212–1216. https://doi.org/10.1109/MIC.2013.6758177

[11] Yi-feng XU, Chun-ming Chen, & Yun-qing XU. (2008). An Improved K-Means Clustering Algorithm [J]. Computer Applications and Software, 25(3), 275–277.

[12] Saputro, D. T., & Sucihermayanti, W. P. (2021). Penerapan Klasterisasi Menggunakan K-Means Untuk Menentukan Tingkat Kesehatan Bayi Dan Balita Di Kabupaten Bengkulu Utara. Jurnal Buana Informatika, 12(2), 146–155. https://doi.org/10.24002/jbi.v12i2.4861

[13] REGINA, S. (2019). Algoritma K-Means Untuk Clustering Kualitas Kinerja Karyawan Pada Pt Clariant Adsorbents Indonesia. Repository.Nusamandiri.Ac.Id. https://repository.nusamandiri.ac.id/index.php/unduh/item/231441/SANDRA-REGINA-11150344.pdf

[14] S. Angra dan S. Ahuja, Analysis of Student's Data Using Rapid Miner, Journal on Today's Ideas – Tommorow's Technologies, 4, 49–58, 2016. https://doi.org/10.15415/jotitt.2016.41004

[15] N.N, "Why RapidMiner," RapidMiner, [Online]. Available: https://rapidminer.com/why-rapidminer/. [Diakses 30 april 2023].

[16] Aprillia, D., Baskoro, D. A., Ambarwati dan, L., Wicaksana, I. W. S. (2013). Belajar Data Mining dengan RapidMiner, Jakarta.

[17] Prayoga, Y., Tambunan, H. S., & Parlina, I. (2019). Penerapan Clustering Pada Laju Inflasi Kota Di Indonesia Dengan Algoritma K-Means. Brahmana: Jurnal Penerapan Kecerdasan Buatan, 1(1), 24–30. https://doi.org/10.30645/brahmana.v1i1.4

[18] Sardi, H. Y., Budayawan, K., New, T., Pendidikan, P., Informatika, T., Teknik, F., & Negeri, U. (2020). Klasifikasi Tingkat Kelulusan Mahasiswa Elektronika Menggunakan Algoritma Naïve Bayes Classifier (Studi Kasus : Pendidikan Teknik Informatika FT-UNP) P – ISSN 8(4), 2302–3295. https://doi.org/10.24036/voteteknika.v8i4.110394

[19] Davies, D. L., Bouldin, D. W. (1979). A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 2, 224. https://doi.org/10.1109/TPAMI.1979.4766909

[20] Gustrianda, R., & Mulyana, D. I. (2022) Penerapan Data Mining Dalam Pe ilihan Produk Unggulan dengan Metode Algoritma K-means Dan K-Medoids, J. Media Inform. Budidarma, 6(1), 27. https://doi.org/10.30865/mib.v6i1.3294

[21] Kouissi, M., En-Naimi, E. M., & Zouhair, A. (2022). Hybrid Approach for Wind Turbines Power Curve Modeling Founded on Multi-Agent System and Two Machine Learning Algorithms, K-Means Method and the K-Nearest Neighbors, in the Retrieve Phase of the Dynamic Case Based Reasoning. International Journal of Online and Biomedical Engineering, 18(6), 110–122. https://doi.org/10.3991/ijoe.v18i06.29565

7 AUTHORS

Dony Novaliendry, Universitas Negeri Padang, Padang, Indonesia.
Tegar Wibowo, Universitas Negeri Padang, Padang, Indonesia.
Noper Ardi, Politeknik Negeri Batam, Batam, Indonesia.
Tiolina Evi, Perbanas Institute, Jakarta, Indonesia.
Dwi Admojo, Perbanas Institute, Jakarta, Indonesia.