Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Application of the K-Means Method for Disease Clustering in Medical Records at Puskesmas Jatiwangi

Firly Maharani Putri[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*
firlymaharani98@gmail.com [1*]

Abstract

Medical record data is often underutilized, limiting opportunities to analyze disease distribution patterns. At Puskesmas Jatiwangi, such data is primarily used for documentation purposes, providing minimal insights to support data-driven decision-making. This study applies the K-Means Clustering method to group disease types based on specific similarities, enabling better identification of disease distribution patterns.
The study utilized 556 patient medical records from October 2024, including attributes such as age, gender, and disease diagnosis. The analysis followed the Knowledge Discovery in Databases (KDD) process, which involves data selection, preprocessing, data transformation, K-Means algorithm implementation, and clustering evaluation using the Davies-Bouldin Index (DBI). Testing was conducted with varying k values from 2 to 10 to determine the optimal number of clusters.
The results indicated that the best DBI value of 0.847 was achieved at k = 2, forming two main clusters. The first cluster represented common diseases such as acute respiratory infections (ARIs) and toothaches, while the second cluster included specific conditions like sciatica and acute lymphadenitis. This study demonstrates that the K-Means method is effective for clustering medical record data, providing valuable insights into disease distribution patterns and aiding in the development of targeted health policies.

Keywords: *K-Means, Medical Records, Clustering, Davies-Bouldin Index, Puskesmas Jatiwangi*

1. Introduction

Public health plays a crucial role in ensuring a good quality of life. The medical records at Puskesmas Jatiwangi document various diseases, making them valuable for monitoring the health status of the local community. However, this data is often underutilized. By clustering disease types using the K-Means method, disease patterns can be analyzed more effectively. This approach is expected to enhance understanding of the characteristics and distribution of diseases in specific areas.

Currently, much of the medical record data at Puskesmas Jatiwangi has not been fully utilized to study the types of diseases prevalent in the region. Disease grouping and identification are still performed manually, which is time-consuming and requires significant attention to detail. This limitation hampers understanding of public health trends and disease distribution patterns. Typically, the available data is only used for documentation, offering little insight into deeper health trends. Moreover, as the volume of data increases, managing and analyzing it becomes increasingly challenging.

The K-Means method is expected to facilitate the automatic clustering of diseases based on features in medical record data. It may also assist in identifying the most common diseases, allowing healthcare programs to prioritize them effectively. The lack of technological application in health data processing has led to incomplete dissemination of information about public health conditions. Thus, further research is necessary to leverage clustering techniques for improving public health data management.

2. Literature Review

2.1. Previous Studies

A study conducted by Rahayu Anggraini highlights that health is a fundamental right for every individual, as significant as education and income. At Puskesmas Ujung Batu in Rokan Hulu Regency, patient medical records are stored solely based on the sequence of diseases, requiring further data processing. This study applied the K-Means algorithm to cluster 3,875 medical record entries using five attributes: gender, participant type, diagnosis, discharge status, and address. The findings showed that the best division resulted in two clusters, with Cluster 1 containing 710 entries and Cluster 2 containing 3,165 entries. A Silhouette Coefficient value of 0.646 indicated good clustering quality. This approach aims to help Puskesmas manage patient data more effectively.[1].

Castaka Agus Sugianto's research on Puskesmas Cigugur Tengah, which serves around 150 patients daily, reveals a continuous increase in data, much of which is archived without further use. The study clustered patient data based on acute and non-acute diseases using the K-Means and K-Medoids algorithms. Results indicated that 93% of patients suffered from acute diseases. The K-Means algorithm outperformed K-Medoids in terms of Davies-Bouldin Index (DBI) values. This research aims to help Puskesmas and health departments understand disease patterns and enhance health education efforts.[2].

Okta Jaya Harmaja's study emphasized that increased public awareness of health and better access to services have led to a rise in patient numbers at Puskesmas. To address this challenge, the K-Means Clustering algorithm was applied to efficiently group disease types. Using RapidMiner 10.1.2, 949 patient records were analyzed based on age, gender, and disease diagnosis, resulting in three clusters: low (198 patients), medium (227 patients), and high (524 patients). The high cluster predominantly consisted of males aged 40 and above, with acute respiratory infections (ARIs) being the most common diagnosis.[3].

M. Agung Vafky Ideal's study explored how patient complaints are responses influenced by internal factors like genetics and external factors like the environment. Understanding these complaints helps Puskesmas prevent diseases more effectively and provide appropriate health services. The study used six months of medical records data covering 72 complaint categories. Data were processed using the K-Means method, which grouped data through partitioning. The classification results assisted Puskesmas in analyzing patient complaint patterns and supported decision-making for disease prevention and health services.[4].

Windania Purba's research addressed the complexity and volume of medical record data in hospitals during the digital era. This data includes patient information, diagnoses, treatments, and medical histories, requiring efficient management to improve healthcare services, decision-making, and medical research. Using the K-Means Clustering algorithm, the study classified medical records at RS Royal Prima Medan. Cluster 1 included 1,827 patients with serious conditions like emergencies, orthopedics, and heart diseases, while Cluster 4 included 417 patients with conditions like urology, ENT, and neurology. The results are expected to aid the hospital in health promotion and disease prevention based on gender and treatment types.[5].

2.2. 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].

2.3. 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], [8], [9].

2.4. 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.[10], [11], [12], [13].

2.5. 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.[14].

2.6. 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.[15].

3. Research Method

3.1. Data Source

This study used secondary data derived from medical records at Puskesmas Jatiwangi, Majalengka Regency, collected in October 2024 and organized in Excel format. The research aimed to cluster disease types using the K-Means method to analyze disease patterns in the region. The data was verified for accuracy and relevance, ensuring only pertinent records were used. The results are expected to contribute to improving healthcare services at Puskesmas Jatiwangi.

3.2. Data Collection Techniques

Data collection involved direct observation and interviews at Puskesmas Jatiwangi. The data consisted of 556 patient records with 15 attributes, extracted from the Puskesmas information system on October 23, 2024. Observations focused on gathering updated medical records from October 2024. The data underwent a structured process, from downloading to dataset preparation for analysis. Additionally, official documents, such as reports and data collection procedures from Puskesmas, were reviewed to ensure the completeness, consistency, and validity of the data used for clustering models, supporting the study's reliability.

3.3. Data Analysis Techniques

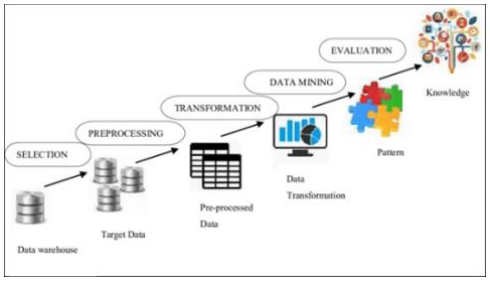

Fig. 1: Stages of the KDD Process

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 9.10.011 tool:

4.1.1. Data Before Preprocessing

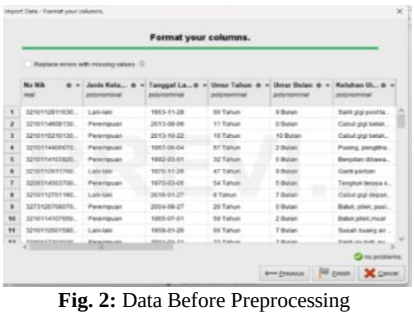

Fig. 2: Data Before Preprocessing

Figure 2 shows the data before preprocessing, obtained from the medical records of Puskesmas Jatiwangi for the period of October 2024.

4.1.2. Determining the Number of Clusters

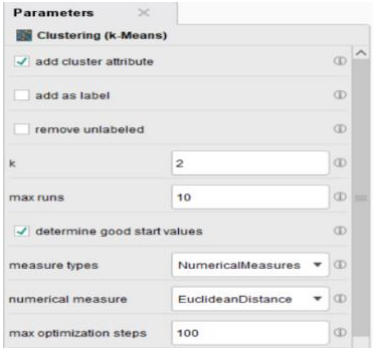

Fig. 3: Determining the Number of Clusters

Figure 3 illustrates the method for determining the value of *k*. The value of *k* is selected based on the smallest Davies-Bouldin Index (DBI). In this dataset, the smallest DBI value is achieved at *k = 2* 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	0.847
3	1.102
4	1.377
5	1.465
6	1.403
7	1.344
8	1.373
9	1.340
10	1.359

Based on Table 1, the lowest value is observed in the cluster with a DBI of 0.847.

4.1.3. Data Preprocessing

Fig. 4: Data Preprocessing

In Figure 4, the data shown has undergone an attribute selection process required for data processing in clustering disease types based on specific characteristics.

4.1.4. Clustering Process in RapidMiner

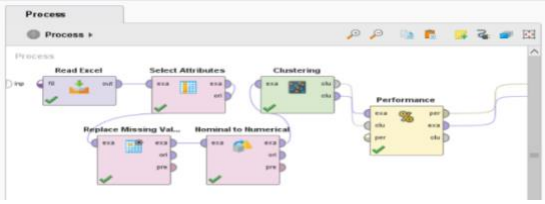

Fig. 5: Clustering Process in RapidMiner

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 *Diagnosis*, *Age (Years)*, *Primary Complaint*, *Height*, *Weight*, *Heart Rate*, *Diastolic Blood Pressure*, *Systolic Blood Pressure*, *Respiratory Rate*, and *Waist Circumference*.
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=2* 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

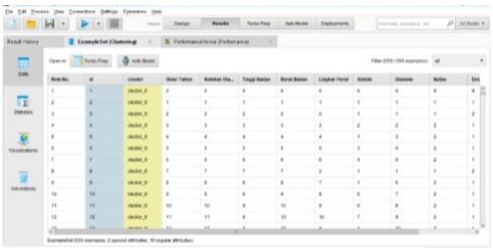

Fig. 6: Clustering Results Data in RapidMiner

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 two (2) clusters with varying numbers of items. Cluster 0 contains 326 items, while Cluster 1 contains 230 items. The total number of items across both clusters is 556.

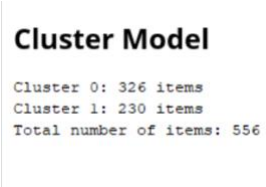

Fig. 7: Cluster Model

4.2.2. Disease Diagnosis Cluster Results

The following are the clustering results from the medical record data, as shown in Figure 8, which displays the cluster results from the application of the K-Means algorithm using the Scatter/Bubble plot type.

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

1. **Cluster 0** consists of 326 data points, predominantly representing acute diseases, particularly upper respiratory tract infections. The most common conditions found in this cluster are: first, *Acute Nasopharyngitis (common cold)* – 47 items; second, *Acute Upper Respiratory Infection, unspecified* – 25 items; third, *Dyspepsia* – 17 items; and fourth, *Acute Lymphadenitis, unspecified* – 4 items. Common symptoms experienced by patients include cough, cold, mild to moderate fever, dizziness, fatigue, sore throat, and nausea. The age range of patients is quite varied, but most are between 1 and 40 years old, indicating that children and young adults are more susceptible to seasonal infections. Diseases in this cluster are generally not severe and can be treated with simple measures such as antipyretics, antihistamines, and adequate rest.

2. **Cluster 1**, with a total of 230 items, is dominated by chronic diseases or conditions that require more complex medical management. The primary diseases in this cluster include: first, *Disturbances in Tooth Eruption* – 51 items; second, *Necrosis of Pulp* – 9 items; third, *Dyspepsia* – 12 items; and fourth, *Myalgia* – 16 cases. Common symptoms reported include toothache, nausea, and abdominal pain due to digestive issues, as well as muscle soreness, particularly in adults and the elderly. Most patients are between 30 and 60 years old, suggesting that chronic conditions like dental problems and digestive disorders are more prevalent in this age group. Diseases in this cluster typically require ongoing medical care, such as dental visits, treatment for digestive disorders, or physiotherapy.

The presence of *Dyspepsia* in both clusters is due to the variation in symptoms and patient characteristics. In **Cluster 0** (17 cases), patients experience acute symptoms such as nausea and heartburn, with an age range of 19–61 years. In contrast, in **Cluster 1** (12 cases), patients

aged 7–68 years exhibit more complex symptoms, including digestive issues, shortness of breath, fever, or body aches. This difference highlights the variability of *Dyspepsia* symptoms, influenced by factors such as blood pressure or body mass index.

5. Conclusion

The research conducted at Puskesmas Jatiwangi, using 556 medical record data, indicates that the optimal value of *K* for the K-means algorithm is 2, with a DBI of 0.847, which suggests a good clustering evaluation. Future research is recommended to use a larger and more diverse dataset to improve accuracy, and to explore other clustering algorithms to compare performance and determine the best method. Combining a broader dataset with different algorithms may provide deeper insights into disease type patterns, thereby enhancing the reliability of the 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.
11. May all the kindness and support extended be rewarded manifold by Allah SWT.

References

[1] R. Anggraini, E. Haerani, J. Jasril, and I. Afrianty, "Pengelompokkan Penyakit Pasien Menggunakan Algoritma K-Means," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 6, p. 1840, 2022, doi: 10.30865/jurikom.v9i6.5145.

[2] C. A. Sugianto, A. H. Rahayu, and A. Gusman, "Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Cigugur Tengah," *Journal of Information Technology*, vol. 2, no. 2, pp. 39–44, 2020, doi: 10.47292/joint.v2i2.30.

[3] O. J. Harmaja, H. Halawa, W. S. Hulu, and S. Loi, "Implementasi Algoritma K-Means Clustering Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Pulo Brayan," *Sains dan Teknologi*, vol. 5, no. 1, pp. 150–157, 2023.

[4] M. A. V. Ideal, "Klasifikasi Keluhan Pasien terhadap Data Rekam Medis Pasien dengan Menggunakan Metode K Means," *Jurnal Sistim Informasi dan Teknologi*, vol. 5, pp. 1–6, 2022, doi: 10.37034/jsisfotek.v5i1.151.

[5] W. Purba, G. A. Sembiring, A. Saputra, T. Turnip, B. Jua, and I. Manihuruk, "Penerapan Data Mining Untuk Pengelolaan Data Rekam Medis Menggunakan Metode K-means Clustering Pada Rumah Sakit Royal Prima Medan," *Jurnal TEKINKOM*, vol. 6, no. 1, pp. 158–168, 2023, doi: 10.37600/tekinkom.v6i1.857.

[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 Mahasiswa Teknik Informatika)*, vol. 7, no. 6, pp. 3363–3369, 2024, doi: 10.36040/jati.v7i6.8197.

[7] S. Rahmayani, S. Sumarno, and Z. A. Siregar, "Analysis of K-Means Algorithm for Clustering of Covid-19 Social Assistance Recipients," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 1, pp. 77–84, 2022, doi: 10.55123/jomlai.v1i1.166.

[8] S. Syahputra, S. Ramadani, and A. M. H. Pardede, "Menentukan Strategi Promosi Menggunakan Algoritma Clustering K-Means," *JOISIE (Journal Of Information Systems And Informatics Engineering)*, vol. 4, no. 1, pp. 7–14, 2020.

[9] S. Suhada and A. M. H. Pardede, "PENGKLASTERAN DOKUMEN DENGAN MENGGUNAKAN ALGORITMA SUPPORT VECTOR CLUSTERING," *JSIK (Jurnal Sistem Informasi Kaputama)*, vol. 3, no. 2, pp. 1–6, 2019.

[10] K. A. Kholil, N. Rahaningsih, and R. Danar Dana, "Penerapan Data Mining Untuk Clustering Penyakit Diare Menggunakan Algoritma K-Means (Studi kasus: Puskesmas Beber)," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 3, pp. 3124–3131, 2024, doi: 10.36040/jati.v8i3.9616.

[11] S. Ramadani, I. Ambarita, and A. M. H. Pardede, "METODE K-MEANS UNTUK PENGELOMPOKAN MASYARAKAT MISKIN DENGAN MENGGUNAKAN JARAK KEDEKATAN MANHATTAN CITY DAN EUCLIDEAN (STUDI KASUS KOTA BINJAI)," *Journal Information System Development (ISD)*, vol. 4, no. 2, pp. 15–29, 2019.

[12] S. Ramadani, I. Ambarita, and A. M. H. Pardede, "METODE K-MEANS UNTUK PENGELOMPOKAN MASYARAKAT MISKIN DENGAN MENGGUNAKAN JARAK KEDEKATAN MANHATTAN CITY DAN EUCLIDEAN (STUDI KASUS KOTA BINJAI)," *Journal Information System Development (ISD)*, vol. 4, no. 2, pp. 15–29, 2019.

[13] N. Damayanti, A. M. H. Pardede, and M. Sihombing, "Pengelompokan Jumlah Produksi Pipa dengan Menggunakan Metode Clustering Berdasarkan Ukuran," *VISA: Journal of Vision and Ideas*, vol. 3, no. 1, pp. 240–250, 2023.

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

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