

IMPLEMENTATION OF DATA MINING FOR INPATIENT PATIENT DATA CLUSTERING USING THE K-MEANS CLUSTERING ALGORITHM

Aniar¹, Niska Salsiani sinta^{2*}, Mariyanto Nurshamsul³

^{1,2,3}Politeknik Baubau, Kota Baubau, Indonesia

ARTICLE INFORMATION

Received: 3 October 2025

Revised: 11 November 2025

Accepted: 15 December 2025

DOI:

KEYWORDS

Keywords : Data Mining, K-Means Clustering, Inpatients, Electronic Medical Records, Wajo Community Health Center.

CORRESPONDING AUTHOR

Name : Niska Salsiani sinta

Address : Jl. Lakarambau. Kel. Lipu. Baubau

E-mail : niskasalsiani@gmail.com

ABSTRACT

The large availability of electronic medical record inpatient data at Wajo Public Health Center has not been optimally utilized and is only used for administrative purposes and routine reporting. The utilization of these data has not been directed toward analyzing patient characteristic patterns to support strategic health service planning. This study aims to implement data mining techniques using the K-Means Clustering algorithm to group inpatient data at Wajo Public Health Center to map patients' demographic and clinical characteristics in a structured manner. The research method used is a quantitative approach with a cluster analysis design. Data were extracted from the EMR system of Wajo Public Health Center in 2026, with a total sample of 1,527 patient records after undergoing preprocessing stages including data cleaning, attribute selection (age and diagnosis), and data transformation using Python programming in Google Colab. The determination of the optimal number of clusters was evaluated using the Elbow Method. The results showed that the most optimal number of clusters is 4 clusters ($K=4$). Cluster 0 (217 patients) is dominated by adults with diverse complaints (abdominal colic, vertigo, partus, ARI). Cluster 1 is the largest group (553 patients) dominated by adults with digestive system disorders (dyspepsia and GEA). Cluster 2 (449 patients) is dominated by infants and children with acute infectious disease patterns (GEA, ARI, febris). Cluster 3 (308 patients) is dominated by adult patients with a combination of acute and chronic/metabolic diseases (dyspepsia, GEA, ARI, diabetes, UTI). In conclusion, the K-Means algorithm is effective in grouping complex medical record data into structured information as a basis for targeted health service planning recommendations.

INTRODUCTION

A Public Health Center (Puskesmas) is a first-level health care facility that provides and coordinates promotive, preventive, curative, rehabilitative, and/or palliative health services within its working area (Permenkes No 19, 2024). In the provision of health services, Puskesmas are required to use Electronic Medical Records (EMR) as part of the health service information system. An Electronic Medical Record is a medical record created using an electronic system for the implementation of medical records and serves as one of the subsystems of a health facility information system that is integrated with other subsystems (Permenkes RI No 24, 2022). The implementation of EMR enables the storage of patient identification, examination, diagnosis, treatment, procedures, and other patient service information in large quantities. Therefore, these data do not only function as documentation and administration, but also have the potential to be utilized as a source of information for improving the quality and planning of health services.

The large volume of inpatient patient data stored electronically creates a need for more systematic data management and analysis so that the information contained within it can be utilized optimally. Inpatient services generate administrative and clinical data, including patient identification, age, sex, primary and secondary diagnoses, length of stay, medical procedures, and medical summaries, which can be used as a source for analyzing patient characteristics (Permenkes RI No 24, 2022). However, based on initial observations conducted at Wajo Public Health Center on Monday, February 23, 2026, through an interview with one of the inpatient registration officers, it was found that inpatient patient data stored in the electronic system was still primarily utilized for administrative purposes and routine reporting. The utilization of these data had not yet been optimized to identify

patient characteristics and disease distribution patterns that could be used as supporting information for health service planning.

The utilization of medical record data to produce more meaningful information can be carried out through data mining, which is a process of processing data using statistical, mathematical, artificial intelligence, and machine learning approaches to extract information and discover relevant patterns from large databases (Adiputra, 2021). The management of patient admission services and medical records is also a crucial component in supporting the smooth delivery of healthcare services. (Sinta & Sulistiawan, 2022). In the health sector, data mining can be used to analyze electronic medical record data to identify patterns of patient characteristics and disease distribution that are not easily observed through manual examination (Wahyu Surya et al, 2023). One method suitable for this purpose is clustering, an unsupervised learning method that groups data based on the degree of similarity in characteristics without using predetermined class labels (Setyaningtyas et al., 2022). The application of clustering can help identify specific structures or patterns in large datasets, thereby facilitating analysis, supporting decision-making, and producing more structured information (Putra et al., 2022).

The K-Means algorithm is one of the clustering methods that works by dividing data into a number of groups based on the proximity of data points to the center point (centroid) of each group (Basalamah & Setyadi, 2023). The K-Means process involves determining the number of clusters, selecting initial centroids, calculating distances using Euclidean Distance, assigning data based on the nearest distance, and updating the centroids iteratively until stable groups are obtained (Basalamah & Setyadi, 2023). The appropriate number of clusters can be determined using the Elbow Method, while cluster separation quality can be evaluated using the Silhouette Score, which ranges from -1 to 1, with values closer to 1 indicating better cluster quality (Putri Khairani Ritonga, 2025). K-Means has the advantages of relatively fast computation, ease of implementation, and suitability for processing large datasets, although it has limitations because it is sensitive to outliers and noise and requires the number of clusters to be determined in advance (Ahmed et al., 2020).

Previous studies have shown that K-Means has potential for identifying patient patterns based on medical record data. Shepyantoni et al. (2024) showed that the application of K-Means to inpatient data of BPJS participants at Kaur Regency Regional General Hospital produced two main groups based on age, sex, BPJS class, and disease diagnosis, with Cluster I dominated by female patients aged 40–59 years with BPJS class III and diseases such as type 2 diabetes, while Cluster II was dominated by male patients aged 20–39 years with diagnoses such as dyspepsia. The study by (Novaliendry et al, 2023) also showed that the application of K-Means to medical record data was able to produce four patient groups based on medical variables, thereby providing an overview of disease distribution patterns to support health service planning. Nevertheless, research on the application of K-Means to inpatient patient data at the Puskesmas level, particularly by utilizing EMR data to map patient characteristics and diagnosis patterns at Wajo Public Health Center in Baubau City, remains limited.

Therefore, this study aims to implement data mining using the K-Means Clustering algorithm on inpatient patient data at Wajo Public Health Center, Baubau City, by analyzing patient characteristics based on age, sex, and diagnosis; conducting data preprocessing and transformation; determining the optimal number of clusters using the Elbow Method and Silhouette Score; and interpreting the diagnostic characteristics of each cluster as information that can support health service planning and medical information management.

METHODS

This study used a quantitative approach with a data mining method using the K-Means algorithm. The study population consisted of all inpatient patient data recorded in the Electronic Medical Record (EMR) of Wajo Public Health Center, Baubau City, in 2024–2025, while the sample consisted of all data meeting the criteria for completeness of age, sex, diagnosis based on ICD-10, length of stay, and type of financing, using a total sampling technique. The study was conducted at Wajo Public Health Center, with data collected through observation, secondary data from the EMR, and literature review.

The data were processed using Microsoft Excel and analyzed with Python through Google Colab. The analysis stages included data selection and preprocessing in the form of data cleaning, removal of incomplete and duplicate data, format adjustment, and selection of relevant attributes. Subsequently, the optimal number of clusters was determined using the Elbow Method and Silhouette Score, followed by distance calculations using Euclidean Distance and the K-Means clustering process until patient groups were formed based on similarity of characteristics.

RESULTS & DISCUSSION

Characteristics of Inpatient Patient Data and Preprocessing

The results showed that after preprocessing the Electronic Medical Record (EMR) data of inpatient patients at Wajo Public Health Center in 2024–2025, a total of 1,527 patient records met the criteria for analysis. The preprocessing stage involved data cleaning, removal of incomplete and duplicate data, format adjustment, and variable transformation so that the data could be used in the clustering process. The variables used included age, sex, disease diagnosis, length of stay, and type of financing, with age and diagnosis serving as the main characteristics in the clustering process. The initial characteristics showed that the adult age group was the most dominant, consisting of 919 patients (60.2%) of the total dataset.

These results indicate that the EMR data of Wajo Public Health Center contained a sufficient number of records for clustering, but still required preparation to ensure accurate analysis. Preprocessing was necessary because incomplete or duplicated medical record data, as well as inconsistent formats, could affect the results of patient grouping. Data transformation was also necessary because the K-Means algorithm performs grouping based on distances between data points, meaning that the variables used must be in a computable form. This is consistent with the characteristics of the K-Means algorithm, which groups data based on their proximity to centroids and can be affected by the characteristics and scale of the data used (Ahmed et al., 2020).

Determination of the Optimal Number of Clusters

The optimal number of clusters was determined before applying the K-Means algorithm using the Elbow Method and Silhouette Score. The Elbow Method results showed that the Within Cluster Sum of Squares (WSS) decreased significantly from K=1 to K=4 and then began to level off, indicating that the elbow point was at K=4. Based on these results, four clusters were selected for the K-Means process because they provided a simpler and more interpretable grouping structure.

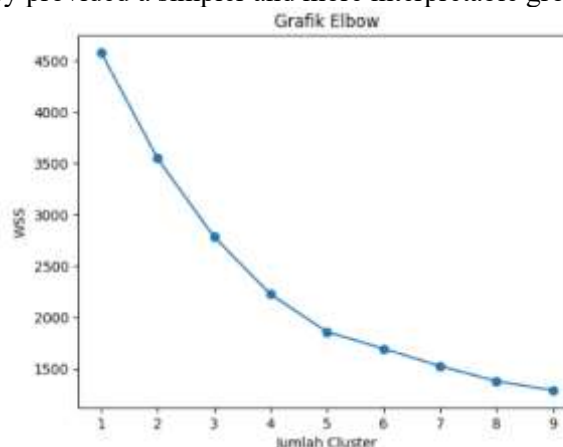


Figure 1. Results of Determining the Number of Clusters Using the Elbow Method

Based on Figure 1, the elbow point can be observed at K=4, indicating that increasing the number of clusters beyond K=4 did not produce a significant decrease in WSS. In addition to the Elbow Method, evaluation was also conducted using the Silhouette Score. The evaluation results showed that the highest Silhouette Score occurred at K=9, with a value of approximately 0.32. However, K=4 was still selected by considering the Elbow Method results, the insignificant increase in Silhouette Score after K=4, and the ease of interpreting the groups for the service needs of Wajo Public Health Center.

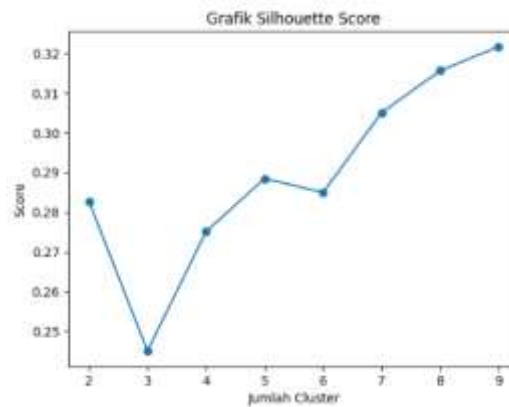


Figure 2. Silhouette Score Evaluation Results

These results show that the determination of the number of clusters in this study considered more than one evaluation indicator. The Elbow Method helps determine the point at which adding clusters no longer produces a substantial decrease in WSS, while the Silhouette Score is used to assess the quality of separation between clusters. According to Zhang & Li (2022) and (Putri Khairani Ritonga, 2025), the Silhouette Score ranges from -1 to 1, with values closer to 1 indicating better cluster separation. The selection of K=4 in this study also considered interpretability, because four patient groups are easier to analyze in the context of age and diagnosis characteristics and can provide a simpler overview for the health service needs of Wajo Public Health Center.

Results of Patient Data Clustering

After the number of clusters was set at four, the clustering process was conducted using the K-Means algorithm with the assistance of Python in Google Colab. The K-Means algorithm groups data based on similarity of characteristics by calculating the distance of each data point to the centroid using Euclidean Distance. The process produced four patient groups, namely Cluster 0, Cluster 1, Cluster 2, and Cluster 3. The number of patients and the main characteristics of each cluster are presented in Table 1.

Cluster	Number	Percentage	Dominant Age	Dominant Diagnosis
0	217	14,2%	Adults (184; 84.8%)	Abdominal colic, vertigo, partus, ARI
1	553	36,2%	Adults (480; 86.8%)	Dyspepsia, GEA
2	449	29,4%	Children (188; 41.9%) and infants	GEA, ARI, febris
3	308	20,2%	Adults (229; 74.4%)	Dyspepsia, GEA, ARI, diabetes, UTI

Based on Table 1, Cluster 1 was the largest group, consisting of 553 patients (36.2%), while Cluster 0 was the smallest group, consisting of 217 patients (14.2%). Cluster 0 was dominated by adult patients, totaling 184 patients or 84.8%, with frequently occurring diagnoses including abdominal colic, vertigo, partus, and ARI. These characteristics indicate that Cluster 0 represents an adult patient group with relatively diverse inpatient cases.

Cluster 1 was the largest group and was dominated by adult patients, totaling 480 patients or 86.8%. The most frequently found diagnoses in this group were dyspepsia, with 326 cases, and GEA, with 108 cases. The dominance of these two diagnoses indicates that digestive system disorders were a prominent characteristic among adult patients in Cluster 1. This finding can provide supporting information for Wajo Public Health Center to strengthen education on dietary patterns, clean and healthy living behaviors (PHBS), diarrhea prevention, and service preparedness related to digestive tract disorders.

Cluster 2 consisted of 449 patients (29.4%) and had different characteristics because it was dominated by infants and children. The most frequently found diagnoses were GEA, ARI, and febris, indicating a pattern in which acute infectious diseases were more dominant among younger age groups. This condition can be considered in strengthening child health services, education on diarrhea and ARI prevention, and monitoring of acute infectious cases among infants and children.

Cluster 3 consisted of 308 patients (20.2%) and was dominated by adult patients with diagnoses of dyspepsia, GEA, ARI, diabetes, and UTI. These characteristics indicate a combination of

acute and chronic or metabolic diseases, particularly diabetes, among adult patients. Therefore, the clustering results for Cluster 3 can provide supporting information for strengthening the management of non-communicable diseases, particularly diabetes, while improving education on the prevention and control of infectious diseases.

Interpretation of Clustering Results for Health Services

The clustering results showed that each group had different age and diagnosis characteristics, providing an overview of inpatient patient patterns at Wajo Public Health Center. Cluster 0, Cluster 1, and Cluster 3 were dominated by adult patients, while Cluster 2 showed a greater representation of infants and children. These differences indicate that the application of K-Means can help transform electronic medical record data into more structured information groups based on similarities in patient and disease characteristics.

These findings are consistent with Shepyantoni et al. (2024), who showed that the application of K-Means to inpatient data of BPJS participants at Kaur Regency Regional General Hospital was able to form two patient groups based on age, sex, BPJS class, and disease diagnosis. The results are also consistent with (Novaliendry et al., 2023), which showed that the application of K-Means to medical record data was able to produce patient groups based on medical variables and provide an overview of disease distribution patterns to support health service planning. Thus, the findings at Wajo Public Health Center strengthen previous research showing that the K-Means algorithm can be used to identify patient group patterns based on medical characteristics.

Overall, the application of data mining using the K-Means algorithm to EMR data of inpatient patients at Wajo Public Health Center produced four patient groups with different characteristics, with Cluster 1 as the largest group and Cluster 2 showing more prominent characteristics of infants and children. These results can be utilized as supporting information to understand disease patterns, determine health education priorities, and assist in planning services based on patient characteristics. Therefore, this study demonstrates that EMR data can be utilized not only for administration and reporting, but also developed as a source of information to support health service decision-making.

CONCLUSION

This study successfully implemented data mining using the K-Means Clustering algorithm on inpatient patient data from the Electronic Medical Record (EMR) of Wajo Public Health Center, Baubau City, for 2024–2025. After preprocessing, 1,527 patient records met the analysis criteria, with the adult age group being dominant, consisting of 919 patients (60.2%). The determination of the number of clusters using the Elbow Method resulted in four clusters, namely Cluster 0 with 217 patients (14.2%), Cluster 1 with 553 patients (36.2%), Cluster 2 with 449 patients (29.4%), and Cluster 3 with 308 patients (20.2%), with different age and diagnosis characteristics in each cluster. Cluster 1 was the largest group and was dominated by adult patients with diagnoses of dyspepsia and GEA, while Cluster 2 was dominated by infants and children with diagnoses of GEA, ARI, and febris. These results indicate that the K-Means algorithm can help identify patterns of patient and diagnosis characteristics in a more structured manner, thereby providing supporting information for health service planning at Wajo Public Health Center.

The limitations of this study include the use of EMR data from 2024–2025 and variables limited to the data available, as well as the determination of the number of clusters depending on the characteristics of the dataset. Therefore, future studies may develop the variables used and compare other clustering algorithms to obtain more comprehensive grouping results.

ACKNOWLEDGEMENTS

The authors would like to thank Wajo Public Health Center, Baubau City, for the opportunity and permission granted to conduct this study. Appreciation is also extended to the medical record officers and related parties who assisted the researchers in obtaining inpatient patient data and provided the information required during the study. The authors also express their gratitude to the academic supervisors who provided guidance, feedback, and direction throughout each stage of the research and the preparation of this article. In addition, the authors thank all parties who provided assistance and support throughout the research process.

REFERENCES

- Adiputra, I. N. M. (2021). CLUSTERING PENYAKIT DBD PADA RUMAH SAKIT DHARMA KERTI MENGGUNAKAN ALGORITMA K-MEANS. *Information System and Emerging Technology Journal*. Vol. 2, No. 2, Desember 2021 99 CLUSTERING, 2(2), 99–105.
- Ahmed, M., Seraj, R., Mohammed, S., & Islam, S. (2020). *The k-means Algorithm : A Comprehensive Survey and Performance Evaluation*. 1–12. <https://doi.org/10.3390/electronics9081295>
- Basalamah, A. T., & Setyadi, R. (2023). Penerapan Algoritma K-Means Clustering Pada Tingkat Penyelesaian Pendidikan Di Provinsi Indonesia. *Jurnal J-COM (Jurnal Informatika Dan Teknologi Komputer)* Vol. 04 No. 02 (2023) 114 – 121 1., 04(02), 114–121. <https://doi.org/DOI:10.55377/j-icom.v4i2.7893>
- Novaliendry, D. (2023). Online and Biomedical Engineering. *International Journal of Online and Biomedical Engineering (IJOE)*, 19(12), 144–155. <https://doi.org/https://doi.org/10.3991/ijoe.v19i12.42147>
- Permenkes No 19. (2024). *Peraturan Menteri Kesehatan Republik Indonesia Nomor 19 Tahun 2024 Tentang Penyelenggaraan Pusat Kesehatan Masyarakat*.
- Permenkes RI No 24. (2022). PERATURAN MENTERI KESEHATAN REPUBLIK INDONESIA NOMOR 24 TAHUN 2022 TENTANG REKAM MEDIS. In *BPK RI* (Vol. 151, Issue 2). BPK RI.
- Putra, A. Z., Pinem, R. W., Silalahi, S., Gulo, F., & Adityo, J. A. (2022). Classification of Covid-19 Patient Spread Rate By Age and Region With K-Means Algorithm. *Jurnal Dan Penelitian Teknik Informatika Volume 6, Number 3, July 2022*, 6(3), 1085–1089. <https://doi.org/DOI:https://doi.org/10.33395/sinkron.v7i3.11603>
- Putri Khairani Ritonga, M. S. H. (2025). Analisis Perbandingan Silhouette dengan Elbow pada Algoritma. *METIK JURNAL VOLUME. 9 NOMOR.1 2025*. <https://doi.org/10.47002/metik.v9i1.1027>
- Setyaningtyas, S., Nugroho, B. I., & Arif, Z. (2022). TINJAUAN PUSTAKA SISTEMATIS PADA DATA MINING : STUDI KASUS ALGORITMA K-MEANS CLUSTERING. *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang - Vol. 10 No. 2 (2022) 52-61*, 10(2), 52–61. <https://doi.org/DOI:https://doi.org/10.21063/jtif.2022.V10.2.52-61>
- Sinta, N. S., & Sulistiawan, W. (2022). Faktor-Faktor yang Menghambat Pelayanan Penerimaan Pasien Rawat Jalan. *Jurnal Sains Dan Kesehatan*, 1(2), 75–87. <https://doi.org/doihttps://doi.org/10.57151/jsika.v1i2.46>
- Wahyu Surya Nanda, Akim M.H. Pardede, M. S. (2023). *Indonesian Journal of Education And Computer Science*. 1(3), 82–88. <https://doi.org/Site:https://jurnal.intekom.id/index.php/indotech>