JMM: Journal of Mechanical and Manufacture e-ISSN: 2829-4424
Vol. 5 No 1, Oktober 2025 , pp. 26-32
DOI: 10.31949/jmm.v5i1.12233

DATA MINING APPROACH TO THE EVALUATION OF THE MOST COMMON DISEASES: K-MEANS CLUSTERING STUDY IN PUSTU LAPANG SEDAR IN 2023

Chindi Fariha Tuzahra[1], Whydiantoro[2]
[1,2]Industrial Engineering, Universitas Majalengka
[1]chindift@gmail.com

Abstract
This study investigates disease patterns at PUSTU Lapang Sedar in 2023 by employing a data mining approach, specifically utilizing the K-Means Clustering method. The primary objective is to identify prevalent diseases and enhance resource allocation for effective health management. Data was collected from health reports, including various disease types and case numbers, and analyzed using the WEKA application. The results revealed three distinct clusters: Dominant Disease (ARI with an average of 661.33 cases), Moderate Illness (Dyspepsia with 253.33 cases), and Rare Diseases (Myalgia with 1631 cases). This clustering not only highlights the most common health issues but also provides a framework for targeted prevention strategies and improved healthcare services. By leveraging technology in data analysis, this research contributes to the growing body of knowledge in public health and supports evidence-based decision-making at the local level.

Keywords: *Data mining; K-Means Clustering; disease evaluation; PUSTU Lapang Sedar; WEKA*

Submitted: 2025-10-01	Revised: 2025-10-10	Accepted: 2025-10-12

Introduction

Public health is one of the fundamental aspects of social and economic development, and a measure of a country's welfare. A health center (Puskesmas) is a functional implementing unit that serves as a center for health development, a center for fostering community participation in health, and a primary health care center that carries out its activities comprehensively, integratively, and sustainably within a community residing in a specific area. (Nurhayati, 2016) Efforts to improve public health cannot be separated from the need to understand and manage health data appropriately. The patient information system at the health center is an information system that handles activities such as queue management, registration, and patient medical records (Putri & Kurniasari, 2020) In Indonesia, basic level health facilities, such as Puskesmas Pembantu (PUSTU), play an important role in health services that directly reach communities in the regions. In this case, Lapang Sedar PUSTU plays a crucial role as a primary health care provider serving the community in Lapang Sedar village, Majalengka. Disease monitoring and evaluation at the PUSTU aims not only to record the number of cases, but also to gain deeper insight into the patterns of frequently occurring diseases. One method of data analysis that can be used to identify patterns of disease spread is data mining. Data mining involves the process of finding patterns and hidden information from large and complex data, which cannot always be identified through conventional analysis methods (Pangestu & Ridwan, 2022) Through a data mining approach, researchers can uncover certain characteristics or patterns that may not be directly apparent. In the context of this research, the K-Means Clustering method was used to explore disease data recorded at Sedar Field Hospital throughout 2023. K-means clustering, as one of the non-hierarchical data clustering methods, partitions existing data into one or more clusters or groups. Data with similar characteristics are grouped into the same cluster, while data with different characteristics are grouped into other clusters. The resulting groups or clusters provide valuable knowledge/information for policymakers in the decision-making process (Nur Khormarudin et al., 2022) K-Means is a method of clustering data based on similar characteristics, which allows the identification of disease groups based on their intensity and frequency of occurrence. The objective of data clustering is to minimize the objective function set during the clustering process, which generally aims to minimize the variation within a cluster and maximize the variation between

clusters (Silitonga & Morina, 2018)**.** This method was implemented using WEKA software, WEKA (Waikato Environment for Knowledge Analysis) is a data mining software developed by the University of Waikato, New Zealand. It was first implemented in 1997 and became open source in 1999 (Mardiana & Nyoto, 2015). WEKA consists of various tools that can be used for tasks such as data pre-processing, classification, regression, clustering, association rules, and visualization (Sharma et al., 2012) WEKA an open-source application that supports data mining analysis with various machine learning algorithms, including K-Means. The legal basis for this research is supported by health regulations that emphasize the importance of systematically collecting and managing health data. Law No. 36/2009 on Health, for example, mandates the government and health institutions to conduct data-based disease control and prevention efforts. Through effective data management, it is expected that health services can be improved sustainably, putting the needs of the community as the top priority. In addition, Minister of Health Regulation No. 2052/2011 on Medical Records requires health facilities to manage health data accurately and comprehensively to support evidence-based decision-making processes. Based on these legal foundations, disease data management at Lapang Sedar PUSTU is not only operationally important, but also essential from a regulatory and health policy perspective.

This research has several urgencies and practical benefits. Firstly, through analyzing disease data using data mining methods, particularly K-Means Clustering, this research can provide a clearer picture of the most prevalent diseases in the working area of PUSTU Lapang Sedar. The clustering results allow the identification of dominant disease groups, which can be the basis for planning more effective allocation of resources, such as medicines and medical equipment. Secondly, by knowing disease patterns and trends, more targeted prevention programs can be designed and implemented. Thirdly, this study can support the improvement of service quality at PUSTU by providing evidence-based data that can be used for better decision-making by health facility management. In this digital era, the use of technology in data analysis is an absolute necessity. The results of this study not only help in understanding disease trends, but can also be used as a reference in making health policies at the local level. By utilizing the WEKA application and the K-Means Clustering method, this research is expected to be the first step in the application of data mining technology in the regional health sector. The results of this study are also expected to be a contribution to the public health literature, especially related to the implementation of data mining in disease analysis, which ultimately helps in strategic planning in the health sector.

Research Method

This study aims to evaluate disease patterns at Sedar Fieldhouse PUSTU in 2023 using a data mining approach, specifically the K-Means Clustering method with the WEKA application. The clustering process is done in three levels based on disease dominance. The following are the stages of the method used:

1. Problem Identification
   The first step was to identify issues related to the high rates of certain diseases at Sedar Fieldhouse PUSTU, which required clustering to understand dominant disease patterns. This analysis is expected to support prioritization of resource allocation and prevention efforts.
2. Literature Study
   A literature study was conducted to understand data mining approaches in analyzing health data, especially the K-Means Clustering method and its application in the WEKA application. The literature studied includes references on health data clustering and disease pattern analysis to improve the quality of health services.
3. Data Collection

The data collected came from health reports at Lapang Sedar PUSTU in 2023. This data included the type of disease, number of cases, age of patients, and other relevant data for cluster analysis. Data collection was done through report observation and access permission from PUSTU, and following research ethics standards.

4. Data Clustering with K-Means
   The data was analyzed using the K-Means Clustering method in the WEKA application, with the following steps:
   - Data Preprocessing: Data is cleaned and prepared to match the WEKA input format. Missing values, outliers, and other irrelevant data are addressed at this stage.
   - Determination of Cluster Parameters: The number of clusters is determined based on the goal of grouping diseases into three main clusters: dominant diseases, moderate diseases, and rare diseases. With this three-level approach, each disease can be grouped based on the number of occurrences.
   - K-Means Clustering Process: The K-Means method was applied in WEKA to cluster diseases based on dominance patterns, resulting in three main disease clusters according to their incidence rates.
5. Data Analysis
   The clustering results from the WEKA application were then analyzed to understand the characteristics of each cluster:

- Moderate Disease Cluster: Identifies diseases with moderate incidence rates that require regular attention and monitoring.
- Dominant Disease Cluster: Identify diseases with the highest number of cases that require priority treatment and prevention.
- Rare Disease Cluster: Identifies low-incidence diseases that still need to be monitored to prevent unexpected spikes.
- Through the stages of this method, this study is expected to provide a clear picture of disease patterns at Lapang Sedar PUSTU. The division into three disease dominance clusters allows for better health prioritization and resource allocation.

The data on the name of the disease and its number in 2023 obtained from the Sedar Field Center is as presented in table 1 below, then the data that has been collected is stored in excel data with the .CSV extension format as shown in figure 1. CSV (Comma-Separated Values) is a simple file format used to store data in tabular form. Each line in a CSV file represents one row of data, and each value (or column) in that row is separated by a comma (,). This format is often used because it is easy to read and write by various applications, including spreadsheets such as Microsoft Excel, Google Sheets, and programming software such as Python.

Table 1. Disease name data of PUSTU Lapang Sedar 2023

No.	Disease Name	Total
1	ARI	1631
2	Dyspepsia	841
3	Myalgia	592
4	Hypertension	551
5	Atritis	394
6	Cepalgia	294
7	Gastritis	249
8	Diarrhea	236
9	Dermatitis	213
10	Febris	134

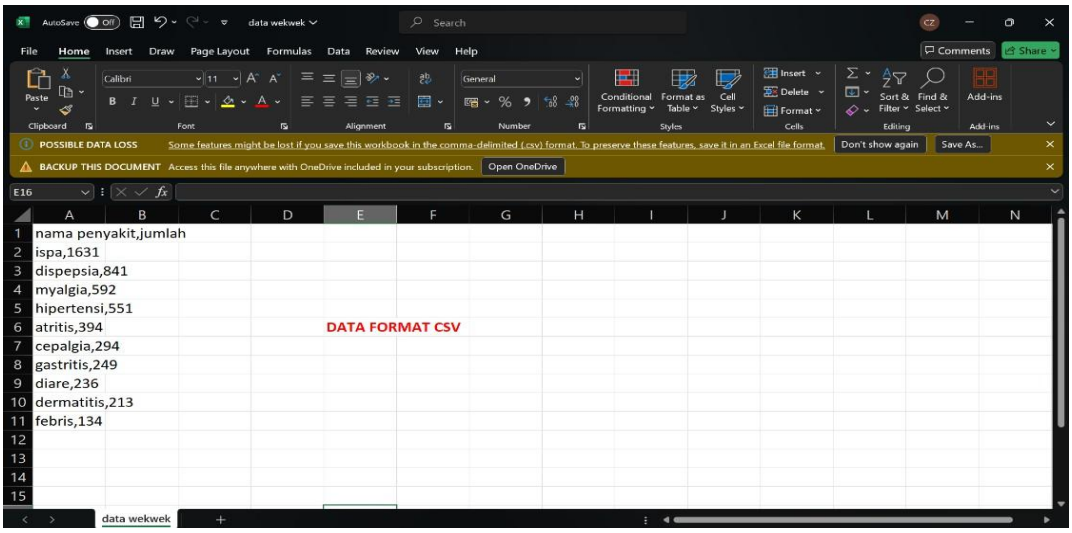


Figure 1. The .CSVs format extension

After the data is stored in a .CSV file, the next step is to determine the attributes that will be used in the clustering process so that it can be applied to the WEKA application, as shown in Figure 2.

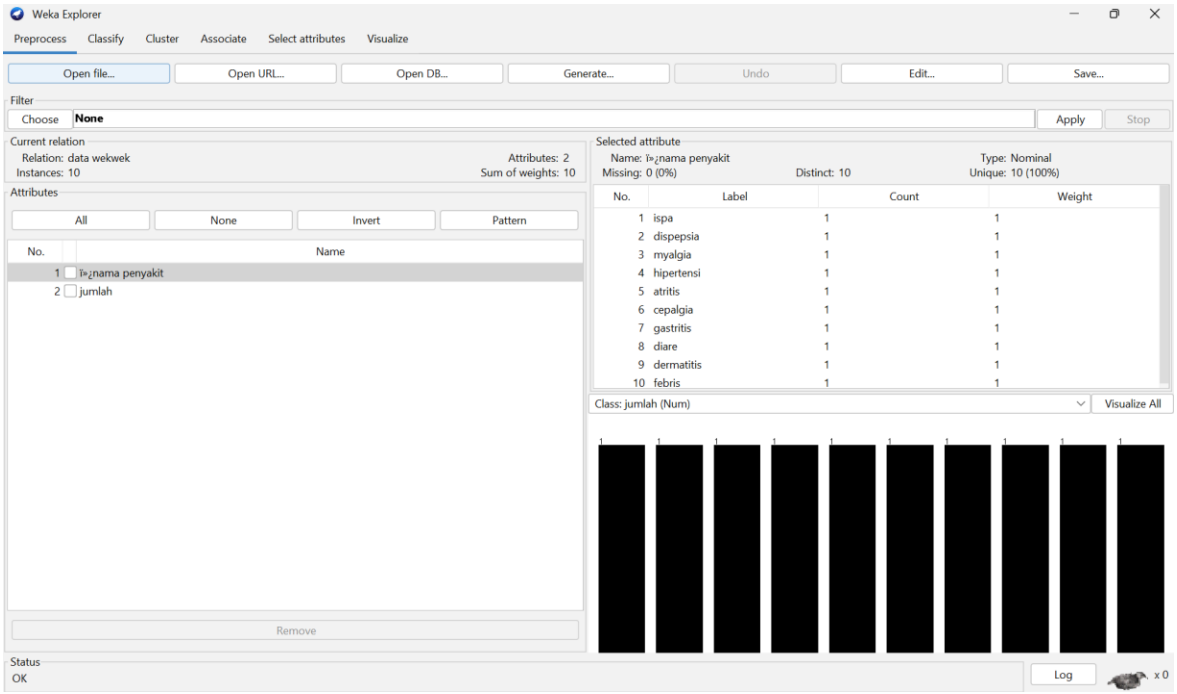
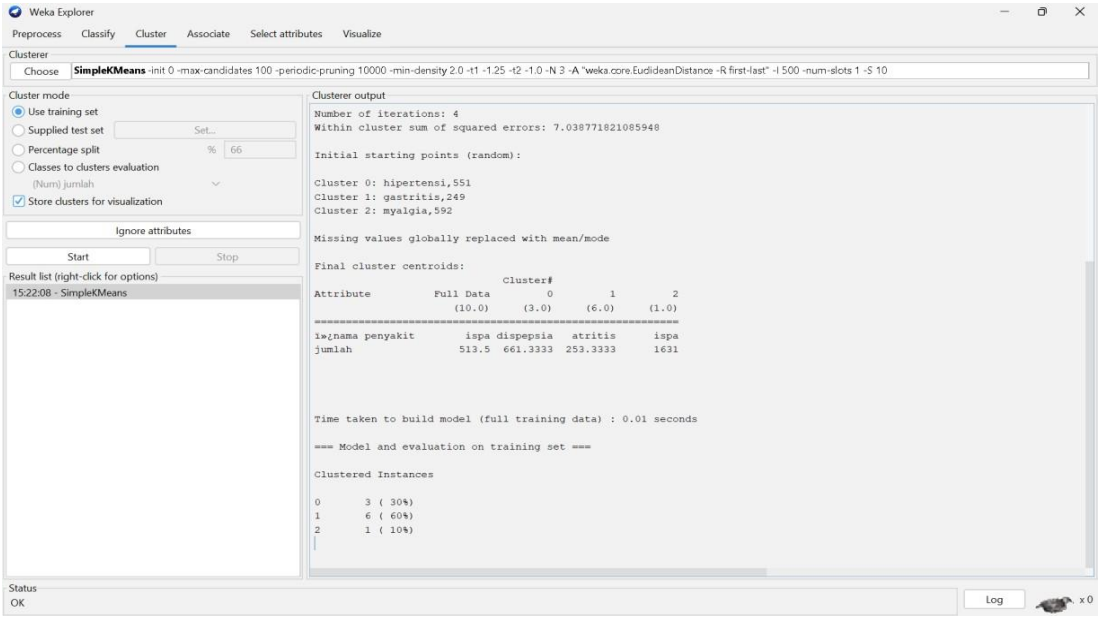


Figure 2. Preprocess WEKA software

The clustering process is done by applying the simple K-Means algorithm, which is one of the methods for clustering. The first step is to determine the amount of data to be clustered through input into the WEKA application. In this research, clustering is done by creating 3 clusters, where the user enters the numCluster value as 3 (three) and presses the Ok button.

After that, click the start button to run the application, then WEKA will start the clustering process and the results obtained are shown in Figure 3 as follows:

Figure 3. Clustering results with WEKA



The explanation for figure 3 is as follows:

1. Number of Iterations: 4

This indicates that the K-Means algorithm has performed the clustering process for 4 iterations. During each iteration, the algorithm updates the cluster centroid positions and evaluates whether the data members have been allocated to the right clusters. Normally, the algorithm will stop if there is no significant change in the cluster allocation or if it reaches the specified maximum iteration limit.

2. Within Cluster Sum of Squared Errors: 7.038771821085948

This value measures how well the data is grouped in the resulting clusters. The lower this number, the better the clusters are at approximating their respective centroids. In this case, a value of 7.04 indicates that there is relatively little variation within each cluster compared to its centroid position.

3. Initial Starting Points (Random)

It indicates the random starting point used by the K-Means algorithm to determine the initial centroid of each cluster:

- Cluster 0: hypertension, 551
- Cluster 1: gastritis, 249
- Cluster 2: myalgia, 592

This starting point can affect the clustering result, and K-Means is sensitive to the choice of starting point, so the use of random points is often necessary to produce a variety of solutions.

4. Missing Values Globally Replaced with Mean/Mode

This indicates that if there is a missing value in the dataset, it is replaced with the mean or mode of the attribute in question. This replacement is important to ensure the data can be used in the analysis without ignoring the missing information.

5. Final Cluster Centroids

This table shows the final centroid for each cluster generated after the iteration is complete. These centroids represent the central characteristics of each cluster and help describe the disease profile within each cluster.

6. Time Taken to Build Model (Full Training Data): 0.01 seconds
The time taken to build the clustering model using the entire training data is 0.01 seconds. This shows the efficiency of the K-Means algorithm in processing data.
7. Clustered Instances
This shows the distribution of the number of agencies in each resulting cluster:

Results and Discussion

The study conducted at PUSTU Lapang Sedar utilized the K-Means Clustering method to analyze disease patterns recorded in 2023. The analysis revealed three distinct clusters of diseases based on their incidence rates, which are critical for understanding health trends and prioritizing resource allocation. The clustering process employed K-Means algorithm iterations, which effectively grouped diseases based on their characteristics. The algorithm completed four iterations, resulting in a Within Cluster Sum of Squared Errors (WCSS) value of approximately 7.04, indicating a good fit for the clusters formed. This low WCSS suggests that the diseases within each cluster are closely related to their respective centroids, reinforcing the validity of the clustering results.
Overall, this study illustrates how data mining techniques like K-Means Clustering can enhance understanding of disease patterns in primary healthcare settings. By identifying dominant, moderate, and rare diseases, healthcare managers at PUSTU Lapang Sedar can make informed decisions regarding resource allocation and health intervention strategies tailored to the specific needs of their community. The findings underscore the potential for data-driven approaches to improve public health outcomes through better planning and targeted interventions.

Cluster 0 = Dominant Disease

identified ARI (Acute Respiratory Infection) as the most prevalent condition, with an average of 661.33 cases. This high incidence indicates that ARI is a significant public health concern in the area, necessitating focused interventions and resource allocation to manage and prevent outbreaks effectively. The data suggests that environmental factors and community health behaviors may contribute to the high rates of ARI, highlighting the need for targeted public health campaigns.

Cluster 1 = Moderate Illness

included dyspepsia, with an average of 253.33 cases. Although not as common as ARI, dyspepsia still represents a considerable health issue that requires ongoing monitoring and treatment. Its moderate incidence suggests that while it may not overwhelm healthcare resources, it is prevalent enough to warrant attention to improve patient outcomes and quality of life.

Cluster 2 = Rare Diseases

encompassed myalgia, which had an average case count of 1631. Despite being categorized as a rare disease, the relatively high number of cases indicates that myalgia still poses a challenge for healthcare providers. This finding emphasizes the importance of recording and managing less frequent conditions to prevent potential increases in incidence and ensure comprehensive healthcare coverage.

Conclusion

1. This study identified ARI as the disease with the highest frequency, which requires priority in resource allocation and health management at Lapang Sedar PUSTU.
2. By using K-Means Clustering, the allocation of resources, such as medicines and medical devices, can be better targeted according to the disease clusters that have been formed.
3. The findings of the study provide a basis for planning more effective prevention programs, particularly for diseases belonging to dominant clusters.
4. The K-Means Clustering-based data mining approach was shown to help improve the quality of health services at PUSTU Lapang Sedar by providing evidence-based data that supports decision-making.

Reference

Pangestu, A., & Ridwan, T. (2022). Penerapan Data Mining Menggunakan Algoritma K-Means Pengelompokan Pelanggan Berdasarkan Kubikasi Air Terjual Menggunakan Weka. *JUST IT : Jurnal Sistem Informasi, Teknologi Informasi Dan Komputer*, *11*(3), 67–71. https://jurnal.umj.ac.id/index.php/just-it/article/view/11591

Silitonga, P., & Morina, I. S. (2018). Klusterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Dengan Menggunakan K-Means Clustering. *Jurnal TIMES*, *6*(2), 22–25. https://doi.org/10.51351/jtm.6.2.2017584

Mardiana, T., & Nyoto, R. D. (2015). Kluster Bag of Word Menggunakan Weka. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *1*(1), 1–5. https://doi.org/10.26418/jp.v1i1.10145

Sharma, R., Alam, M. A., & Rani, A. (2012, August). K-means clustering in spatial data mining using weka interface. In *International conference on advances in communication and computing technologies (ICACACT)* (Vol. 26, p. 30).

Putri, F. P., & Kurniasari, F. (2020). Sistem Informasi Layanan Puskesmas Berbasis Web. *Ultimatics : Jurnal Teknik Informatika*, *11*(2), 89–93. https://doi.org/10.31937/ti.v11i2.1457

Nurhayati, M. (2016). Peran Tenaga Medis Dalam Pelayanan Kesehatan Masyarakat Di Puskesmas Pembantu Linggang Amer Kecamatan Linggang Bigung Kabupaten Kutai Barat. *EJournal Llmu Administrasi Negara*, *4*(1), 2127–2140. https://ejournal.ap.fisip-unmul.ac.id/site/wp-content/uploads/2016/02/Isi Jurnal (02-17-16-12-35-21).pdf

Asroni., Andrian, R. 2015. Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Strudi Kasus Pada Jurusan Teknik Informatika UMM Magelang. Jurnal Ilmiah Semesta Teknika, 18(1), 76-82.

Nur Khormarudin, A., Kevin Synagogue Panjaitan, O. C., Septianingsih, A., Faisal, M., Utami, W. S., Novia Ningsi, L., Satria Tambunan, H., Zai, C., & Komputer, T. (2022). Teknik Data Mining: Algoritma K-Means Clustering. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *1*(2), 116–123. https://djournals.com/klik%0Ahttps://ilmukomputer.org/category/datamining/

Werdiningsih, I., Nuqoba, B. & Muhammadun. (2020). Data Mining Menggunakan Android, Weka, dan SPSS. Surabaya: Airlangga University Press.

Kulkarni, E. G., & Kulkarni, R. B. (2016). Weka powerful tool in data mining. *International Journal of Computer Applications*, *975*, 8887.