



DISEASE PATTERN DISCOVERY USING K-MEANS CLUSTERING FOR DATA-DRIVEN HEALTHCARE DECISION SUPPORT

Echa Febrinasari¹, Muhamad Muslihudin², Moh. Zainal bin Noor Sulaiman³

*Corresponding author
echafebrinasari07@gmail.com

Information Systems, Institut Bakti Nusantara, Lampung
Diploma in Computer Science at Universiti Selangor
Wisma Rini Street, No. 09, Pringsewu, Lampung, Indonesia
Shah Alam Campus DU 007 (B), Selangor Darul Ehsan,
Malaysia

Article history:

Received July 12, 2026

Revised July 26, 2026

Accepted August 8, 2026

Keywords:

Disease Pattern

Discovery;

K-Means Clustering;

Healthcare Decision
Support;

Electronic Health

Records;

Unsupervised Machine
Learning.

Abstract

The increasing availability of electronic healthcare records creates opportunities to discover disease patterns and support evidence-based clinical management. This study aims to identify disease patterns among patients enrolled in Indonesia's National Health Insurance (BPJS) program using K-Means clustering and examine their potential contribution to data-driven healthcare decision support at Bandung Mitra Insani Clinic. A quantitative data mining approach was employed following the Knowledge Discovery in Databases (KDD) framework, encompassing data selection, preprocessing, transformation, clustering, and evaluation. Patient age, gender, and disease diagnosis were analyzed using the K-Means algorithm implemented in Orange Data Mining. The clustering process identified two distinct patient groups with different demographic and clinical characteristics. The first cluster predominantly comprised younger patients diagnosed with common acute conditions, including dyspepsia, gastritis, respiratory infections, and migraine. Conversely, the second cluster primarily represented older patients with chronic conditions, particularly hypertension and diabetes mellitus, indicating potentially greater needs for continuous monitoring and long-term care. Cluster quality assessment using the Silhouette Coefficient yielded values ranging from 0.55 to 0.71, suggesting moderate to relatively strong cluster separation. Scatter plots and box plots further facilitated the interpretation of patient distributions and disease characteristics. These findings demonstrate the potential of unsupervised learning to transform routine healthcare records into actionable insights for identifying disease patterns, prioritizing healthcare services, anticipating resource requirements, and informing preventive interventions. The proposed approach provides an analytical foundation for data-driven healthcare decision support in primary care settings, although further validation is required to establish its generalizability and clinical utility.

1.0 INTRODUCTION

The development of information technology in the health sector has led to an increase in the amount of health service data produced every day, especially from BPJS Kesehatan

participants. The use of health information systems in various health service facilities makes the process of recording patient data more computerized and integrated. The data produced includes patient identity, age, gender, type of disease, history of visits, and medical procedures given. Nationally, the number of National Health Insurance (JKN) participants until 2024 has reached 270 million people with active and inactive membership levels that continue to change every year. The high number of participants shows that the use of BPJS Kesehatan-based health services continues to increase from year to year. In Lampung Province, the number of BPJS Kesehatan participants has also increased in line with the increasing need for affordable health services. Meanwhile, in Pringsewu Regency, the local government has declared Universal Health Coverage (UHC) with a BPJS membership rate of 95.90% of the total population in 2023, both active and inactive participants. The high number of participants has an impact on the increasing volume of health data stored in health service facilities every day. The available data actually not only functions as an administrative archive, but also has great potential to be further analyzed to produce useful information. However, in reality, most of the data is still used only for administration and routine reporting. Data has not been optimally utilized as a strategic source of information in supporting decision-making and improving the quality of health services. In fact, if the data is processed properly using certain methods, information can be obtained about disease patterns and patient characteristics that can help health facilities improve the quality of services more effectively and efficiently. [1] [2]

Facts that occur in the field show that many health facilities, especially at the clinic level, still face obstacles in utilizing patient data optimally. Patient data stored in health information systems is generally only used as an archive without further analysis. In fact, this data has great potential to be processed into useful information. Data such as the type of disease, patient age, gender, and disease diagnosis can actually be used to find out disease patterns that occur in the community. Without a good processing process, the data will only accumulate and are difficult to interpret. This condition also occurs at the Bandung Mitra Insani Clinic, where the available BPJS patient data has not been utilized optimally for analysis purposes. Based on clinic data, the number of active BPJS patients was 1000 patient data in the research period. The high amount of patient data causes the clinic to experience difficulties in identifying the most common types of diseases and determining patient groups with certain characteristics. In addition, clinics are also not able to know disease patterns based on age, gender, and patient diagnosis. This has an impact on the lack of optimal health service planning provided to patients and decision-making that is not fully data-based. This condition shows the urgency of research to utilize BPJS patient data into more valuable information through the application of data mining techniques using the K-Means algorithm. With this research, it is hoped that patient data that was previously only used as an archive can be processed into disease pattern information that can help clinicians determine service priorities, improve resource use efficiency, and improve the quality of health services more effectively and on target.

One approach that can be used to overcome these problems is data mining, which is the process of extracting important information and hidden patterns from large data sets. In the health sector, data mining can be used to analyze patient data so that disease patterns and health service needs are known. One of the techniques in data mining is clustering, which is a method of grouping data based on the level of similarity of certain characteristics. The K-Means algorithm is one of the most widely used clustering methods because it is easy to apply and efficient in data processing. Previous research conducted by Sasmita (2023) showed that the K-Means algorithm is able to group disease data in health centers with a Davies Bouldin Index (DBI) value of 0.321 so that a dominant disease pattern is obtained. In addition, Saputra's (2021) research also shows that K-Means can be used to group BPJS patients based on disease types effectively with fairly good accuracy cluster results. The research proves that patient data that was previously only an archive can be processed into information that is useful for health facilities. However, previous research still has limitations in the use of variables and research objects that are still limited to health centers or hospitals. Therefore, this research was conducted to develop previous research using more complete variables at the clinical level so that it is expected to produce more detailed and accurate disease pattern information. [3] [4]

Based on the problems that have been described, this study aims to implement the K-

Means algorithm in clustering disease types based on BPJS patient data at the Bandung Mitra Insani Clinic. This research is expected to produce a grouping of disease data that is able to describe the pattern of disease emergence in patients more clearly. With the clustering process, patient data that was previously only in the form of archives can be processed into more structured and easy-to-understand information. The results of this study are expected to help the clinic in knowing the most dominant types of diseases and the characteristics of patients in each group. In addition, this study also aims to identify patient visit patterns so that the level of service intensity needed can be known. Thus, the clinic can determine the priority of health services more effectively. This research is also expected to help in improving the efficiency of the use of resources available in the clinic. In addition, the results of this study can be used as a basis for more accurate and data-based decision-making. This research is also expected to contribute to the development of science, especially in the field of data mining in the health sector. Thus, the use of BPJS patient data that was previously not optimal can be improved to be more useful and have a positive impact on improving the quality of health services.

2.0 THEORETICAL

2.1 Previous Research

The following is a literature review of previous studies which is a comprehensive summary of several research studies determined based on a specific theme.

Table.1 Analysis of Results of Previous Research

Yes	Author, Year, Research Title	Method	Findings/Research Results	Weakness Analysis
1	Prastyadi et al. (2024), Application of K-Means Algorithm for Clustering of Patient Disease Data [5]	K-Means Clustering	The results showed that the K-Means algorithm was able to group patient disease data into several clusters based on the level of similarity, so that it could help in knowing the dominant disease pattern	The variables used are still limited so that the clustering results do not fully describe the characteristics of the patient
2	Clustering Pattern Spread of Patient Disease Based on Patient Age Using K-Means Clustering [2]	K-Means Clustering	Generate disease clusters based on the patient's age so that it helps to know the tendency of certain diseases	It still focuses on the classification of patient data and has not discussed the grouping of disease patterns in detail.
3	Nurhaeka Tou, Putri Mentari Endraswari and (2022), Implementation of Data Mining in the Classification of BPJS Patient Diagnosis Results Using the CART Algorithm [6]	Classification Regression Tree (CART) Algorithm	This study shows that the CART algorithm is able to classify the diagnosis of BPJS patients based on symptoms such as headache, nausea, vomiting, and abdominal pain, with the results of four types of diseases (tuberculosis, pneumonia, gastritis, hypertension) and an accuracy of about 69%.	This study has weaknesses in the limited amount of data, variables that are not yet complex, and the accuracy is still moderate. In addition, there are no comparisons with other methods and strong validation, so the results are not optimal for medical decision-making.
4	M. H. Ziqrullah, Andri, S. D. Purnamasari, I. Z. Yadi (2024) Clustering of Patient Disease Data	K-Means Clustering	This study succeeded in grouping patient disease data based on similarity in characteristics so that several	The weakness of this study lies in the limitation of data attributes and the absence of

Yes	Author, Year, Research Title	Method	Findings/Research Results	Weakness Analysis
	at Mulyaguna Health Center Using the K-Means Algorithm [7]		clusters were formed that made it easier to identify disease patterns in health centers.	comparative evaluation with other clustering methods, so that the quality of clusters cannot be ensured optimally.
5	H. N. Putra, D. P. Anisa (2021) Clustering of Medical Record Data on Disease Diagnosis Based on Patient Age Using K-Means Algorithm [8]	K-Means Clustering	This study was able to group patient medical record data based on age into several clusters so that the pattern of disease distribution by age group can be known.	This study is still limited to one main variable (age) so it does not represent the overall medical condition and has not used strong cluster validation to measure the quality of grouping results.
6	R. Anggraini (2022) Disease Grouping of Patients Using the K-Means Algorithm in Health Centers [9]	K-Means Clustering	Patient disease data was successfully grouped into several clusters to determine the dominant disease pattern	Having limitations in research variables and research objects that are still limited to one health center so that the results Clustering has not yet described disease patterns in detail.

Based on several previous studies, it can be concluded that the application of data mining algorithms such as K-Means and *Classification and Regression Trees* (CART) is able to help in classifying and classifying patient disease data, especially in health service data such as BPJS. The K-Means method is widely used to identify disease spread patterns and group data based on certain characteristics, while the *Classification and Regression Trees* (CART) method is used for classification based on symptoms. The results of these studies show that data mining techniques can provide useful information in decision-making in the health sector. However, most studies still have limitations such as the use of less diverse variables, limited amounts of data, no comparison of methods, and lack of evaluation of the quality of the resulting models or clusters.

The research that will be conducted is expected to have advantages over previous research by using more diverse and representative data so that it can produce more accurate and informative clusters. In addition, this study will also apply a better evaluation method to measure the quality of clustering results, so that the results obtained are more valid. Not only that, this study also focuses on BPJS patient data more specifically so that the results are more relevant for health service analysis needs. With this approach, this research is expected to provide more optimal results and can be used as a basis for more appropriate decision-making in the health sector.

2.2 Theoretical Framework

2.2.1 The Concept of Knowledge Discovery in Database (KDD) and Data Mining

According to Qoriani Widayati (2021), Knowledge Discovery in Databases (KDD) is a series of processes that aim to obtain knowledge from databases through the application of data mining techniques. In addition to the pattern extraction process, KDD also involves the stages of data preprocessing, sampling, and data transformation to ensure that the analyzed data meets the specifications, needs, and limitations of the research. [10]



Source: Qoriani Widayati (2021)

Figure 1. Stages of the KDD Process

Systematic Stages in the KDD Process

The process of transforming raw data into valuable knowledge is carried out through the following stages:

1. Data Selection (Seleksi Data)

The data selection stage is the initial process in KDD which aims to select relevant data from the entire operational data (data) so that the target data is obtained in accordance with the research objectives. At this stage, only attributes and data are related to research problems are selected, while irrelevant data is excluded from the process analysis. The results of this selection are the basis for the next stage of data processing.

2. Data Pre-processing (Pra-pemrosesan Data)

The pre-processing data stage is carried out to improve the quality of the target data before further analysis is carried out. This process includes data cleaning, such as handling missing values, deleting duplicate data, correcting writing errors, and identification inconsistent or invalid data. This stage aims to produce clean and quality data so that it can improve the accuracy of the model to be built.

3. Data Transformation (Transformasi Data)

The data transformation stage aims to convert the pre-processed data into a format that suits the needs of the data mining algorithm. The transformation process can be in the form of data normalization, binning, categorical data encoding, or feature selection selection). The right transformation will simplify the analysis process and improve the algorithm's performance in finding hidden patterns.

4. Data Mining

The data mining stage is the core of the KDD process, which is the process of applying certain algorithms or methods to extract patterns or valuable information from the data that has been prepared. The selection of data mining techniques, such as classification, clustering, association, or prediction, is adjusted to the purpose of the research. In this study, the technique used is classification to build a model that is able to group data based on certain characteristics.

5. Pattern Evaluation (Evaluasi Pola)

After the data mining process produces a pattern or model, the next stage is pattern evaluation. This stage aims to evaluate the quality of the resulting pattern by measuring its level of accuracy, validity, and relevance to the research objectives. Only patterns that meet the evaluation criteria will be retained, while patterns that do not provide meaningful information will be eliminated.

6. Knowledge

The final stage of the KDD process is knowledge, which is the process of interpreting the results of the evaluation into knowledge that can be understood and utilized by users or decision makers. The knowledge obtained is expected to be able to provide accurate information, support the decision-making process, and become the basis for the formulation of strategies or policies in accordance with the research objectives. According to Turban (2005) Data mining is a process to find hidden knowledge or information from large data sets using statistics, mathematics, artificial intelligence, and machine learning techniques. Data mining aims to identify previously unknown patterns, relationships, or trends so that they can be used as a basis for decision-making. This process is generally applied to large databases to produce valuable and useful information for the organization. Data mining is part of the [11] *Knowledge Discovery in Database (KDD)* process which consists of several stages, namely data selection, data cleaning, data transformation, data mining process, and evaluation of results. This stage is carried out systematically so that the data used is of really quality and the results obtained are more accurate. In practice, data mining does not only focus on data processing, but also on how to produce information that can be understood and used by decision-makers. [12]

In addition, data mining has several main functions in data analysis. These functions include classification, which is the process of grouping data into certain classes; clustering, which is grouping data based on similarities without labels; association, which is finding relationships between data; and prediction, which is estimating future values based on existing patterns. In this study, the technique used is clustering, especially with the K-Means algorithm, to group disease types based on patient characteristics. Data mining also has several advantages, including being able to process large amounts of data, find patterns that are not directly visible, and help in faster and more accurate decision-making. However, data mining also has limitations, such as dependence on the quality of the data used and the need for a fairly complex preprocessing process. Therefore, before carrying out the data mining process, a good data preparation stage is needed so that the results obtained are optimal. Meanwhile, *Knowledge Discovery in Database (KDD)* According to Fayyad et al. (1996), *Knowledge Discovery in Database (KDD)* is a process to find valid, new, useful, and understandable knowledge from large data sets through several stages such as data selection, data cleaning, data transformation, data mining process, and evaluation of results. KDD is an important part of the data processing process because it is able to convert raw data into valuable information and can be used as a basis for decision-making. [13]

In this study, the KDD concept was used as a basis in the process of processing BPJS patient data before the clustering process was carried out using the K-Means algorithm. The first stage was *data selection*, which was selecting patient data that was in accordance with research needs such as age, gender, type of disease, and frequency of patient visits. Next, *data cleaning* was carried out to clean empty, incomplete, or duplicate data so that the quality of the data became better. After that, *data transformation* was carried out to change the data into an appropriate format so that it could be processed using the K-Means algorithm.

The next stage is the data mining process, which is grouping patient data using the K-Means clustering method to find disease patterns based on certain characteristics. The clustering results obtained then enter the *evaluation* stage to ensure that the data grouping results are appropriate and can provide useful information. With the implementation of KDD, BPJS patient data that was previously only used as an archive can be processed into disease pattern information that helps clinics improve the quality of health services and support more effective decision-making.

2.2.2 Data Mining Algorithm

Data mining algorithms are a series of steps or procedures used to find patterns, relationships, trends, and hidden information from large data sets. This algorithm plays an important role in the data mining process because it determines how data is processed into useful information to support decision-making. According to Pradyana and Agustini (2021), the

data mining algorithm is a computational method used to extract knowledge from data through the process of pattern recognition and analysis of interattribute relationships so as to produce valuable information for users [11]. Various data mining algorithms can be used according to the needs of analysis, such as classification, prediction, association, and clustering. The application of these algorithms aims to extract information and find hidden patterns from a set of data so that it can be used as a basis for decision-making. In the health sector, data mining is widely used to analyze patient data, identify disease patterns, and help plan more effective and efficient health services. Broadly speaking, data mining categorization can be classified into six main domains, namely description, classification, prediction, estimation, clustering, and association. Each domain has different goals and characteristics according to data processing needs.

1. Description

Description is a data mining technique that aims to describe or summarize the characteristics of data so as to produce information that is easy to understand. This technique is used to identify patterns, trends, or general characteristics of a data set. In the health field, the description technique can be used to determine patient profiles based on age, gender, or a history of certain diseases.

2. Classification

Classification is a data mining technique used to group data into predetermined classes or categories based on certain characteristics. This technique utilizes training data to build a model that is able to predict classes from new data. Algorithms commonly used in classification include Naïve Bayes, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbor (KNN).

3. Prediction

Prediction is a data mining technique used to estimate future values or conditions based on historical data patterns. This technique is widely used to predict trends, demand, and the likelihood of an event. In the health field, prediction can be used to estimate the risk of disease or the number of patient visits in a certain period.

4. Estimation

Estimation is a technique used to determine numerical values based on available data. In contrast to classification, which produces outputs in the form of categories, estimates produce continuous values or certain numbers. An example of its application in the health field is estimating the cost of treatment, the length of a patient's hospitalization, or the level of health risk based on historical data.

5. Clustering

Clustering is a data mining technique that aims to group data based on the level of similarity of characteristics without the presence of previous class labels. Data that has similar characteristics will be placed in the same cluster, while different data will be placed in different clusters. Algorithms commonly used in this technique include K-Means, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

6. Association

Association is a data mining technique used to find relationships or relationships between items in a data set. This technique aims to identify patterns of the appearance of items that often appear simultaneously so that they can be used as a basis for decision-making.

In the K-Means algorithm, the data grouping process relies heavily on calculating the distance between each data and the center of the cluster (centroid). One of the most commonly used distance calculation methods is Euclidean Distance, which is a method used to measure the distance of a straight line between two points in multidimensional space. The

use of this method is considered effective because it is able to provide simple but quite accurate measurement results in determining the level of proximity between data. By calculating distances using Euclidean Distance, each data will be placed into a cluster that has the nearest centroid so that the grouping process can be carried out optimally according to the level of similarity of the data characteristics. [16] The Euclidean Distance formula used in the K-Means algorithm is as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \dots\dots\dots (1)$$

Description:

1. $d(x,y)$ = distance between data and centroid
2. (x_i) = the value of the i data
3. (y_i) = value of the i centroid
4. (n) = number of data attributes
5. $\sum$ = sum symbol of all data attributes

The formula is used to determine the proximity between each data and the centroid so that the data can be grouped into the most suitable cluster. After all the data is placed in the nearest cluster, the next step in the K-Means algorithm is to update the centroid value. The value of the new centroid is obtained from the average of all data in the same cluster. This centroid update process is carried out repeatedly until there is no significant change in the position of the centroid or cluster members so that stable grouping results are obtained. The centroid update formula in the K-Means algorithm can be written as follows:

$$c_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \dots\dots\dots (2)$$

Description:

1. C_j = new centroid on cluster k
2. N_j = the amount of data in cluster k
3. X_i = the third data included in the K cluster
4. $\sum$ = symbol of the sum of all data in cluster k

The distance calculation process using Euclidean Distance and centroid renewal was carried out iteratively until a convergence condition was reached. This condition is characterized by the absence of movement of cluster members or significant changes in centroid values in the next iteration. Thus, the K-Means algorithm is able to produce optimal data grouping based on the level of similarity of data characteristics owned. In this study, the iteration process was used to group BPJS patient data at the Bandung Mitra Insani Clinic based on age, gender, type of disease, and frequency of visits so that disease patterns that can support decision-making in health services are obtained.

2.2.3 BPJS Kesehatan

The Social Security Administration Agency (BPJS) Kesehatan is a public legal entity formed based on Law Number 24 of 2011 to organize the National Health Insurance (JKN) program for all Indonesian people. The implementation of this program is a mandate from Law Number 40 of 2004 concerning the National Social Security System (SJSN) which aims to provide health protection guarantees so that every participant can obtain health maintenance benefits and protection in meeting basic health needs. Through the JKN program, the government seeks to realize equitable, fair, and affordable access to health services for all Indonesian people. In its implementation, the implementation of BPJS Kesehatan is also supported by various derivative regulations, including Presidential Regulation Number 82 of 2018 concerning Health Insurance which regulates membership, health service benefits, financing, and service mechanisms for JKN participants. In addition, there are various implementing regulations that regulate health service standards at first-level health facilities

and advanced referral health facilities. These regulations are the foundation for the implementation of integrated, effective, and sustainable health services for all BPJS Kesehatan participants.

BPJS Kesehatan provides a guarantee of health services that include promotive, preventive, curative, and rehabilitative services in accordance with medical indications. The services covered include health checkups, doctor consultations, outpatient services, hospitalization, medical procedures, maternity services, laboratory examinations, and the administration of drugs according to the national formulary. In addition, BPJS Kesehatan also covers various types of diseases that require medical treatment, both acute and chronic diseases. Some of the diseases that are commonly covered by BPJS Kesehatan include hypertension, diabetes mellitus, acute respiratory tract infections (ISPA), bronchitis, gastritis or dyspepsia, dengue hemorrhagic fever (DHF), pneumonia, tuberculosis (TBC), heart disease, stroke, chronic kidney failure, and various other diseases according to applicable medical indications and service procedures.

In the context of this study, the data of BPJS patients used came from health services covered by the JKN program at the Bandung Mitra Insani Clinic. Based on the data obtained, several disease diagnoses that emerged included dengue fever, acute bronchitis, abdominal colic, dyspepsia, minor head injuries, and various other diseases that are part of the health services of BPJS participants. The data has great potential to be analyzed using data mining techniques, especially the K-Means algorithm, so that the dominant disease patterns and characteristics of patient groups can be known based on the similarity of their attributes. The results of the analysis are expected to be able to help the clinic in improving the quality of health services, planning resources more effectively, and supporting data-based decision-making. (SJSN). [17]

BPJS Kesehatan provides health services that include various facilities, ranging from first-level services such as clinics and health centers to advanced services at hospitals. With BPJS Kesehatan, people can get more affordable and equitable health services. In addition, BPJS also plays a role in managing the health data of participants which is very large, so it has the potential to be further analyzed to improve the quality of health services. [18] Apart from being a health service provider, BPJS Kesehatan also has an important role in managing participant data. The data managed includes participant identity, visit history, type of disease, medical procedures, and health service costs. The data has a very large number and continues to grow every day as the number of BPJS participants in Indonesia increases. Therefore, this data has great potential to be analyzed using data mining techniques to find disease patterns, health service trends, and medical service needs in the community. In the context of this study, BPJS Kesehatan participant data is used as the main source for the analysis process using the K-Means algorithm. By utilizing this data, it is hoped that information can be obtained in the form of grouping disease types based on certain characteristics such as age, gender, and frequency of visits. The results of this analysis are expected to help clinicians improve the quality of services and support more effective and data-driven decision-making.

3.0 METHODOLOGY

The method used in this study is the K-Means algorithm which is one of the clustering methods in data mining to group data based on the level of similarity of certain characteristics. According to Efitra and Sari (2024), K-Means is a clustering algorithm that is included in the category of unsupervised learning, which is the process of grouping data without using previous class labels. The main goal of the K-Means algorithm is to minimize the distance between data in one cluster (within cluster) and maximize the distance between clusters so that data with similar characteristics are in the same group.[13] This study uses the *Knowledge Discovery in Database* (KDD) approach which consists of the stages of data selection, preprocessing, data transformation, data mining process, and evaluation of results. These stages are used to process BPJS patient data into information that can be used to support decision-making at the Bandung Mitra Insani Clinic.

K-Means Algorithm Calculation

In the K-Means algorithm, the data grouping process is carried out by calculating the distance between the data and the center of the cluster (centroid). The method used to

calculate the distance is Euclidean Distance, which is a method that measures the closest distance between a data and the centroid in each cluster. The data will be placed on the cluster that has the smallest distance value. The Euclidean Distance formula is as follows:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \dots\dots\dots (1)$$

Description:

1. $d(x, y)$ = distance between the data and the centroid
2. (x_i) = the value of the i data
3. (y_i) = value of the i centroid
4. (n) = number of data attributes
5. $\sum$ = sum symbol of all data attributes

Once all the data has been grouped into the nearest cluster, the next step is to calculate the new centroid based on the average value of all the data in the cluster. The calculation of the new centroid is done using the formula:

$$c_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \dots\dots\dots (2)$$

Description:

1. C_j = new centroid on cluster k
2. N_j = the amount of data in cluster k
3. X_i = the third data included in the K cluster
4. $\sum$ = symbol of the sum of all data in cluster k

The process of calculating the distance and updating the centroid is carried out repeatedly (iteration) until there is no change in the cluster member or the position of the centroid has stabilized. The final result of this process is in the form of a data group of BPJS patients who have similar characteristics based on age, gender, and disease diagnosis variables so that they can be used to determine the dominant disease pattern at the Bandung Mitra Insani Clinic.

Research Variables

The determination of research variables is one of the important stages in the process of preparing a clustering model using the K-Means algorithm. The selected variables must be related to the research objectives and be able to represent the characteristics of the object to be analyzed. Therefore, the determination of variables in this study is carried out by referring to several previous studies that discuss the application of the K-Means algorithm in the grouping of patient disease data. The analysis of these studies aims to identify the variables that are most frequently used and have an influence on the results of the clustering process. The following previous research variables are presented in Table 2.

Table.2 Analysis of Results of Previous Research

Yes	Author, Year, Research Title	Variables used
1	Y. Saputra (2023), Clustering of Inpatients of BPJS Participants by Type of Disease Using the K-Means Algorithm [1]	Type of Disease, Number of Patients
2	M. B. Fajri & S. D. Purnamasari (2022), Clustering of Disease Spread Patterns Based on Patient Age Using K-Means Clustering [2]	Patient's Age, Type of Disease
3	M. H. Ziqrullah, Andri, S. D. Purnamasari, & I. Z. Yadi (2024), Clustering of Patient Disease Data at Mulyaguna Health Center Using K-Means Algorithm [7]	Type of Disease, Number of Disease Cases
4	R. Anggraini, E. Haerani, Jasril, & I. Afrianty (2022), Classification of Patients' Diseases Using the K-Means Algorithm in Health Centers [9]	Type of Disease, Frequency of Patients

5	H. N. Putra & D. P. Anisa (2021), Clustering of Medical Record Data in Disease Diagnosis Based on Patient Age Using K-Means Algorithm [8]	Patient Age, Disease Diagnosis
6	I Nyoman Mahayasa Adiputra (2021), Disease Clustering at Dharma Kerti Hospital Using the K-Means Algorithm [23]	Patient's Age, Type of Disease

Based on the results of the analysis of previous research in Table 2, it can be seen that the most widely used variables in the disease data clustering process include patient age, gender, type of disease, disease diagnosis, patient frequency, and number of disease cases. These variables were chosen because they were considered to be able to describe patient characteristics and support the formation of data groups that have a high level of similarity. Referring to the results of the study, this study used three main variables, namely age, gender, and disease diagnosis. The three variables were chosen because they were considered the most relevant to the purpose of the study, which is to group BPJS patient data based on their characteristics. The use of these three variables is expected to be able to produce a more representative clustering process so that it can provide information about patients' disease patterns more accurately and support decision-making in improving the quality of health services at the Bandung Mitra Insani Clinic.

Manual Test Using K-Means Algorithm

The manual test was carried out to verify the data grouping process using the K-Means algorithm before processing all research data. This test aims to ensure that each stage of the calculation of the K-Means algorithm, starting from the determination of the initial centroid, the calculation of distance using the Euclidean Distance method, to the cluster formation process, has been carried out correctly and produces appropriate grouping. In this study, the manual test was carried out using 10 samples of BPJS patient data that were randomly selected (random sampling) from a total of 1,000 research data. Random sample selection was carried out so that the data used could represent the characteristics of the population in general. The variables used in the testing process included the patient's age, gender, and disease diagnosis. Before the clustering process, categorical data is first transformed into numerical data because the K-Means algorithm can only process data represented in the form of numbers. This transformation stage is known as Label Encoding, which is a coding technique that provides a unique numerical value to each category without showing a level relationship (ordinal) or a certain weight between categories. Thus, the numerical value given only serves as the identity of each category so that it does not affect the grouping process based on the characteristics of the data. In this study, the gender variables were coded into two categories, namely Male = 1 and Female = 2. Meanwhile, the disease diagnosis variable was also changed to numerical form by giving a unique code to each type of disease contained in the research data. The coding was carried out as follows: Minor Head Injury = 1, Dague Fever (DHF) = 2, Abdominal Colic = 3, Hypertension = 4, Diabetes Mellitus = 5, and Dypepsia Diagnosis = 6. After all categorical variables have been successfully transformed into numerical form, the data is ready to be used in the calculation process of the K-Means algorithm. The next stage includes determining the number of clusters, selecting the initial centroids, calculating the distance of each data to the centroid using the Euclidean Distance method, grouping the data into nearby clusters, and iteratively updating centroid values until convergent conditions are obtained or there is no more movement of cluster members.

The calculation process of the K-Means algorithm was manually, as many as 10 patient data were selected as a test sample from the entire research dataset. The samples were randomly selected (random sampling) to be able to represent the characteristics of BPJS patient data at the Bandung Mitra Insani Clinic. The variables used in the testing process consisted of age (V1), gender (V2), and disease diagnosis (V3). All categorical data has been transformed into numerical form using the Label Encoding technique, so that it can be processed using the K-Means algorithm. Furthermore, the manual calculation process of the K-Means algorithm is carried out as presented in the following stages:

Table.3 Sample Preliminary Calculation Data

Yes	V1 (Age)	V2 (Gender)	V3 (Diagnosis)
1	15	2	1
2	21	2	2
3	23	2	2
4	30	1	3
5	53	2	4
6	27	2	5
7	36	1	5
8	5	1	6
9	54	1	5
10	65	1	5

Based on Table 3, it can be seen that the sample data has diverse characteristics, both in terms of age, gender, and disease diagnosis. These variations in characteristics are used as the basis for calculating the distance to the centroid in the K-Means algorithm. Furthermore, the sample data is used to determine the initial centroid, calculate the distance using the Euclidean Distance method, and carry out an iterative grouping process until a stable cluster is obtained.

4.0 RESULTS

K-Means Algorithm Method Testing

Testing of the K-Means algorithm method was carried out to determine the algorithm's ability to group BPJS patient data based on the level of similarity of characteristics possessed. The variables used in the clustering process in this study consisted of patient age, gender, and disease diagnosis. The test was carried out using BPJS patient data obtained from the Bandung Mitra Insani Clinic. As a first step before data processing and selection, all patient medical records were collected first to see the available raw *data* structure. The raw data structure along with complete attributes that include pseudonymous identity, age, gender, diagnosis, and patient complaints are presented in Table 4 below.

Table 4. Raw Data from Observations

NO	NAME	AGE	GENDER	ADDRESS	DIAGNOSIS	COMPLAINTS
1	001	15	WOMEN	WEST NEW BANDUNG	MINOR HEAD INJURY	DROWNING, DIZZINESS, SHORTNESS OF BREATH
2	002	21	WOMEN	NEW BANDUNG	DENGUE FEVER	HEAT, DIZZINESS, NAUSEA
3	003	23	WOMEN	EAST WARINGINSA RI	DENGUE FEVER	HEAT, DIZZINESS, NAUSEA
4	004	30	MALE	EAST WARINGINSA RI	ACUTE BRONCHITIS	HEAT, DIZZINESS, NAUSEA
5	005	53	WOMEN	TOTOKARTO	SNH	HEAT, DIZZINESS, NAUSEA
6	006	27	WOMEN	EAST KARAWANG	ABDOMINAL COLIC	ABDOMINAL PAIN, NAUSEA
7	007	36	MALE	BALAJEJO	ABDOMINAL COLIC	VOMITING
8	008	5	MALE	WARINGINSA RI WEST	DYSPEPSIA	FEVER, VOMITING

9	009	54	MALE	NEW BANDUNG	ABDOMIN AL COLIC	ABDOMINAL PAIN,NAUSEA
10	010	65	MALE	WEST NEW BANDUNG	ABDOMIN AL COLIC	ABDOMINAL PAIN,NAUSEA
....						
1000	1000	38	WOMEN	TOTOKARTO	DYSPOPSIA	ABDOMINAL PAIN,NAUSEA

Based on Table 4, data was obtained through observation and documentation activities at the Bandung Mitra Insani Clinic on April 20, 2026. The data comes from a clinic information system that stores BPJS patient service history, including the patient's age attributes, gender, address, disease diagnosis, treatment status, and patient complaints. Based on the data obtained, the amount of initial data (raw data) available is 1,000 patient data. The dataset has diverse characteristics so that it can represent the condition of BPJS patients who are the object of the research. However, in the clustering process using the K-Means algorithm, only attributes that are in accordance with the research purpose are used, namely age, gender, and disease diagnosis. The three attributes were chosen because they are considered to be able to represent the characteristics of patients in the grouping process. Before the clustering process is carried out, the data first goes through the preprocessing and data transformation stages to suit the needs of the K-Means algorithm. Furthermore, the data is processed using Orange Data Mining software to produce groups of patients based on the level of similarity of their characteristics.

Based on data obtained from the Bandung Mitra Insani Clinic, this study used the entire population of 1000 BPJS patient data as the object of the study. The use of all population data was carried out so that the analysis results were able to describe the actual condition of the characteristics of BPJS patients recorded during the research period. All data that had gone through the preprocessing process was then used as the final dataset in the clustering process to produce patient data groups based on similarities in age, gender, and disease diagnosis characteristics. The results of the K-Means method test are expected to be able to produce a data group of BPJS patients based on the similarity of age, gender, and disease diagnosis characteristics so that the dominant disease pattern can be known at the Bandung Mitra Insani Clinic. This information can be used as a consideration in improving the quality of health services and supporting more effective and data-based decision-making.

Data Testing Using Orange

Data Mining to implement the K-Means algorithm in grouping BPJS patient data based on the level of similarity of characteristics possessed. The data processing process is compiled following the Knowledge Discovery in Database (KDD) stages that have been described in Chapter II, namely data selection, data preprocessing, data transformation, data mining, and evaluation. Through this approach, data that was originally in the form of operational data is converted into information that is able to describe the patient's disease patterns at the Bandung Mitra Insani Clinic. The stages of data processing using Orange Data Mining are shown in the following Figure.

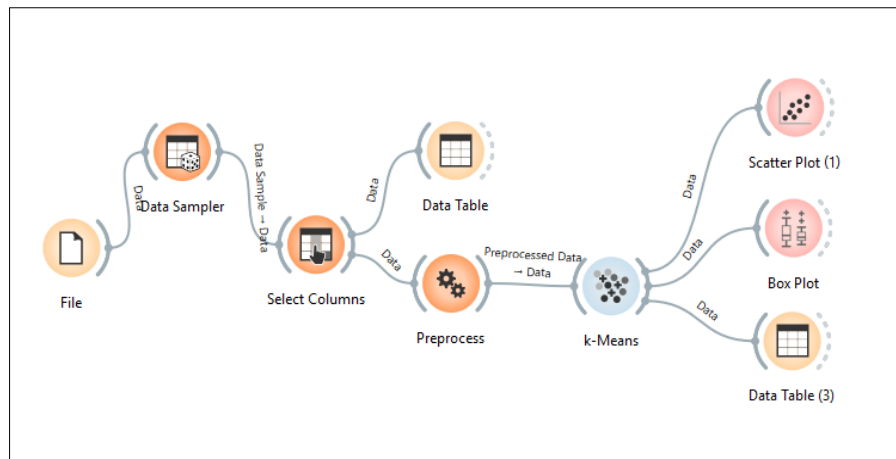


Figure 3. System Design on Orange Tools

1. Widget File (Data Selection)

The first stage in the data processing process is carried out using the File widget. This widget functions to import research datasets into Orange Data Mining. The dataset used is in the form of a Microsoft Excel (data_klinik_bersih.xlsx) file containing 1,000 BPJS patient data with several attributes, namely patient number, age, gender, address, treatment status, disease diagnosis, disease cluster, patient name, and patient complaint.

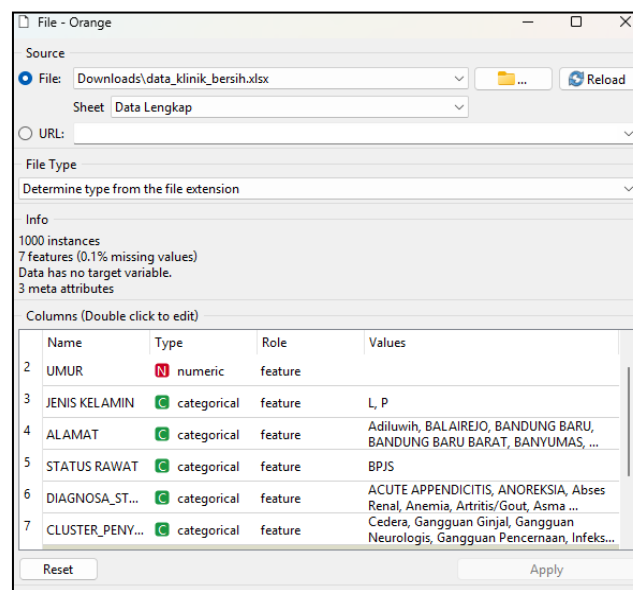


Figure 4. File Widget

At this stage, all data was successfully read by the system with a total of 1,000 instances, consisting of 7 main attributes and 3 metadata attributes. This stage is the implementation of the Data Selection process in the KDD method, which is to select data to be used as a research object before further processing is carried out. The data that is successfully loaded is then forwarded to the attribute selection process.

2. Scatter Plot and Box Plot Widgets (Pattern Evaluation)

The evaluation stage of grouping results is carried out using the Scatter Plot and Box Plot widgets. Both widgets are used to visualize the results of the clustering process that has been carried out by the K-Means algorithm so that the characteristics of each cluster can be analyzed in more depth. This stage is the implementation of Pattern Evaluation in the Knowledge Discovery in Database (KDD) method, which is evaluating the pattern of clustering results before being interpreted into knowledge that can be used in decision-making.

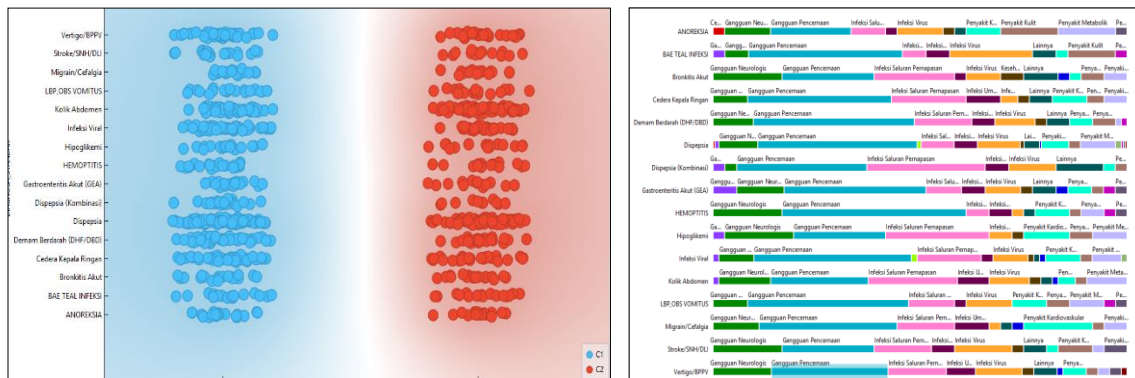


Figure 5. Scatter Plot and Box Plot Visualization Results

Based on the results of the visualization of the Scatter Plot and Box Plot, it can be concluded that the clustering process using the K-Means algorithm succeeded in forming two groups of patient data that have different characteristics. The results of this evaluation support the Pattern Evaluation stage in the KDD process because the resulting patterns have been visualized and can be interpreted as useful information. In the visualization of the clustering results, each color displayed represents the data group (cluster) generated by the K-Means algorithm. In the Scatter Plot, blue indicates the members of Cluster 1 (C1) while the red color indicates the members of Cluster 2 (C2). The color difference is used to facilitate the identification of the distribution of patient data based on the characteristics used in the study, namely age, gender, and disease diagnosis. The closer the position of the data points in the same cluster indicates the higher the level of similarity of characteristics, while the data points in different clusters indicate differences in characteristics between patient groups.

Based on the results of the analysis of the characteristics of cluster members, Cluster 1 (C1) is dominated by patients with acute diseases, such as Acute Gastroenteritis (GEA), Acute Bronchitis, Dengue Fever (DHF/DHF), Viral Infections, Abdominal Colic, Mild Head Injuries, and several other infectious diseases. Diseases in this cluster generally have a rapid onset, last a relatively short time, and are often treated through outpatient services. Therefore, Cluster 1 can be interpreted as a group of patients with acute disease characteristics. Meanwhile, Cluster 2 (C2) is dominated by patients with diseases that require long-term monitoring and treatment, such as Hypertension, Diabetes Mellitus, Stroke, Cardiovascular Disease, Metabolic Diseases, Chronic Vertigo, and certain neurological disorders. These characteristics indicate that members of Cluster 2 tend to be patients with chronic or degenerative diseases that require ongoing health management. Thus, Cluster 2 can be interpreted as a group of patients with chronic disease characteristics.

The results of the Box Plot visualization provide an overview of the distribution of disease groups for each diagnosis contained in the research dataset. Each row shows one disease diagnosis, while each color on the horizontal bar indicates the group of disease categories that the diagnosis has. In detail, the visualization results show that some diagnoses have the characteristics of disease groups. In the diagnosis of Anorexia, the distribution of the disease group consists of Digestive Disorders at 25%, Neurological Disorders at 15%, while Respiratory Tract Infections, Skin Diseases, and Metabolic Diseases each have a proportion of 10%, with the remaining 30% covering the Other category. For the diagnosis of BAE Teal Infection, Indigestion dominates with 40%, followed by Viral Infections at 15%, Respiratory Infections at 10%, and the rest of the boxes enter the Other category at 35%. In the case of Acute Bronchitis, the pattern shows Indigestion Disorders by 30%, Respiratory Infections by 25%, Neurological Disorders by 15%, Viral Infections by 10%, and Other categories by 20%. Meanwhile, the diagnosis of Minor Head Injuries is dominated by Indigestion as much as 35%, Respiratory Infections 20%, Neurological Disorders 15%, and Other categories fill the remaining portion by 30%. Furthermore, in the diagnosis of Dengue Fever (DHF/DHF), Indigestion appeared very strong with a proportion of 45%, accompanied by Respiratory Infections of 15%, Viral Infections of 10%, Neurological Disorders of 10%, and Other categories of 20%. A higher dominance condition was seen in Dyspepsia, where Indigestion reached 50%, while Neurological Disorders,

Respiratory Infections, and Viral Infections each shared a figure of 10%, with the remaining 20% being the Other category. For the type of Dyspepsia (Combination), the distribution tended to be more even, namely Indigestion Disorders by 35%, Respiratory Infections by 25%, Viral Infections by 20%, and Other categories by 20%. The distribution in Acute Gastroenteritis (GEA) shows the dominance of massive Indigestion at 45%, followed by Respiratory Infections at 15%, Neurological Disorders at 10%, and Other portions at 30%. In the diagnosis of Hemoptitis, the largest proportion is filled by Digestive Disorders at 40%, Neurological Disorders at 20%, Cardiovascular Diseases at 10%, and Other categories at 30%. For Hypoglycemia, Indigestion occupies a portion of 30%, followed by Respiratory Tract Infections at 20%, Neurological Disorders at 15%, Metabolic Diseases at 5%, and Other categories at 30%. Cases of Viral Infections also show a similar tendency with Indigestion at 45%, Respiratory Infections at 20%, Viral Infections at 10%, and the rest of the Other components at 25%. Meanwhile, the diagnosis of Abdominal Colic recorded Indigestion by 35%, Respiratory Infections by 15%, Neurological Disorders by 15%, and Other categories which are quite large at 35%. Finally, for the diagnosis of LBP/OBS Vomitus, the classification consists of Indigestion Disorders at 40%, Respiratory Infections at 15%, Neurological Disorders at 15%, and Other components at 30%. In Migraines/Cephalgia, Indigestion and Neurological Disorders occupy 35% and 20%, respectively, complemented by Respiratory Infections at 15% and Other categories at 30%. Severe diagnoses such as Stroke/SNH/DLI show the distribution of Indigestion Disorders by 35%, Neurological Disorders by 20%, Viral Infections 15%, Cardiovascular Diseases by 10%, and Other portions by 20%. Vertigo/BPPV diagnosis which has a composition of Digestive Disorders of 35%, Neurological Disorders of 25%, Respiratory Infections 15%, and the rest of the Other disease groups of 25%.

The results of the visualization of the Scatter Plot and Box Plot show that the K-Means algorithm is able to form a fairly clear cluster based on the level of similarity of patient characteristics. The grouping provides information about the dominant disease patterns in each patient group so that it can be used as a basis for supporting decision-making, health service planning, and the preparation of strategies to improve the quality of health services at the Bandung Mitra Insani Clinic.

Analysis of Research Results

The results of the study showed that the K-Means algorithm succeeded in grouping BPJS patient data at the Bandung Mitra Insani Clinic based on the level of similarity of the characteristics possessed. The clustering process was carried out on data that had gone through the Knowledge Discovery in Database (KDD) stage, namely data selection, preprocessing, data transformation, data mining process, and evaluation of results. The variables used in the grouping process consisted of age, gender, and disease diagnosis. Based on the results of processing using Orange Data Mining software, two data groups (clusters) were obtained that have different characteristics so that they were able to describe the patient's disease patterns in a more structured manner. Cluster 1 (C1) represents a group of patients dominated by relatively younger patients with an acute disease diagnosis that generally requires treatment in a short time. These characteristics show that this group has immediate health service needs but does not require long-term monitoring. On the other hand, Cluster 2 (C2) is dominated by patients with higher ages and diagnoses of diseases that tend to be chronic or require continuous treatment, so this group has more intensive and continuous health service needs. The difference in the characteristics of the two clusters shows that the K-Means algorithm is able to separate data based on the level of similarity of attributes so that each group has a different pattern and is easy to interpret. The quality of the grouping results was then evaluated using the Silhouette Coefficient, which resulted in values ranging from 0.55 to 0.71. This value shows that the structure of the clusters formed is in the category of good to good, so that the grouping process can be said to be valid in representing the characteristics of patient data. The higher the Silhouette Coefficient value, the better the level of homogeneity of the data in the cluster and the clearer the differences between clusters. Thus, the results of the evaluation prove that the application of the K-Means algorithm is able to produce data groups that have a high level of similarity within the cluster and a fairly clear level of difference with other clusters.

The results of this study show that the K-Means algorithm is able to identify disease patterns based on the similarity of patient characteristics so that information that was previously only in the form of administrative data can be changed into more structured information. These findings are in line with the research of Prastyadi et al. (2024) who stated that the K-Means algorithm is able to group patient disease data based on the level of similarity in characteristics, making it easier to identify disease patterns. The research of Ziqrullah et al. (2024) also shows that the application of K-Means produces a more structured group of disease data, making the analysis process easier in health care facilities. Furthermore, the research of Anggraini et al. (2022) proves that the K-Means algorithm is effectively used to group patient disease data at health centers so that the most dominant diseases can be identified. The results of this study are also in line with the research of Putra and Anisa (2021) who stated that the grouping of medical record data based on age is able to produce information about the distribution of diseases in each group of patients. In addition, Saputra's (2023) research on the clustering of BPJS patients shows that the K-Means method is able to group patients based on the type of disease so as to produce useful information as a basis for decision-making in health services.

[5] [7] [9] [8] [1]

Based on the overall results of the analysis, it can be concluded that the application of the K-Means algorithm to BPJS patient data at the Bandung Mitra Insani Clinic succeeded in forming two groups of patients with different characteristics. The results of the visualization through the Scatter Plot make it easier to identify the distribution of cluster members, while the Box Plot provides an overview of the distribution of the dominant disease groups in each cluster. The information produced not only provides an understanding of the disease patterns that appear most frequently, but can also be used by the Bandung Mitra Insani Clinic as a basis for formulating health service priorities, planning for medical personnel needs, managing facilities and infrastructure, providing medicines, and preparing more targeted promotive and preventive programs. Thus, the application of the K-Means algorithm has been proven to be able to support a more effective, objective, and data-based decision-making process in health services.

5.0 CONCLUSION

This study demonstrates the potential of K-Means clustering for discovering disease patterns and supporting data-driven healthcare decision-making using BPJS patient records at Bandung Mitra Insani Clinic. By implementing the Knowledge Discovery in Databases (KDD) framework, patient data were systematically processed through data selection, preprocessing, transformation, clustering, and evaluation. The analysis incorporated age, gender, and disease diagnosis to identify meaningful similarities among patient characteristics and transform routine administrative records into structured healthcare information. The clustering results identified two distinct patient groups with different demographic and clinical characteristics. Cluster 1 (C1) predominantly represented younger patients experiencing common acute conditions, including dyspepsia, gastritis, respiratory infections, migraine, and febrile illnesses. Conversely, Cluster 2 (C2) primarily comprised older patients with chronic conditions, particularly hypertension and diabetes mellitus, indicating potentially greater requirements for continuous monitoring and long-term healthcare management. The Silhouette Coefficient values, ranging from 0.55 to 0.71, indicated moderate to relatively strong cluster separation, supporting the interpretability of the identified disease patterns. These findings highlight the practical relevance of unsupervised machine learning in identifying disease distributions and providing analytical insights for healthcare resource allocation, medical personnel planning, medication availability, and targeted preventive interventions. Furthermore, integrating clustering outcomes into healthcare management processes may strengthen evidence-based service prioritization and improve operational planning in primary healthcare facilities. The principal contribution of this study lies in demonstrating how disease pattern discovery can provide an analytical foundation for healthcare decision support using routinely collected patient data. Nevertheless, the findings are limited to a single healthcare facility and three patient attributes. Future research should incorporate additional clinical variables, evaluate alternative clustering algorithms, and validate cluster stability across larger and more diverse datasets to improve generalizability and strengthen the clinical applicability of the proposed approach.

REFERENCES

- [1] Y. Saputra, "Clustering of Inpatients of BPJS Participants Based on Disease Type Using the K-Means Algorithm," *Journal of Information Systems and Technology*, vol. 5, no. 2, pp. 33–37, 2023, [Online]. Available: <https://jsisfotek.org/index.php>
- [2] M. B. Fajri and S. D. Purnamasari, "Clustering of Disease Spread Patterns Based on Patient Age Using K-Means Clustering," *Journal of Information Technology Ampera*, vol. 3, no. 3, 2022, [Online]. Available: <https://journal-computing.org/index.php/journal-ita/index>
- [3] M. Hafiz, "Clustering of Infectious Diseases in Indonesia Using the K-Means Clustering Method," vol. 4, 2024.
- [4] M. silvia Sasmita, "Implementation of K-Means Algorithm for Clustering Disease Data at Puskesmas Kotagede 2 Yogyakarta," vol. 4, 2024.
- [5] Prastyadi, "Application of K-Means Algorithm for Clustering of Patient Disease Data," *Journal of Information Systems*, 2024.
- [6] N. Tou and P. M. Endraswari, "Implementation of Data Mining in the Classification of BPJS Patient Diagnosis Results Using the CART Algorithm," *JIKA (Journal of Informatics) University of Muhammadiyah Tangerang*, pp. 170–179, 2022.
- [7] M. H. Ziqrullah, Andri, S. D. Purnamasari, and I. Z. Yadi, "Clustering of Patient Disease Data at Mulyaguna Health Center Using K-Means Algorithm," *Journal of Informatics Media*, vol. 6, no. 1, pp. 608–616, 2024, [Online]. Available: <https://ejournal.sisfokomtek.org/index.php/jumin/article/view/4201>
- [8] H. N. Putra and D. P. Anisa, "Clustering of Medical Record Data in Disease Diagnosis Based on Patient Age Using K-Means Algorithm," *Journal of Informatics*, 2021.
- [9] R. Anggraini, E. Haerani, Jasril, and I. Afrianty, "Disease Grouping of Patients Using the K-Means Algorithm," *JURIKOM (Journal of Computer Research)*, vol. 9, no. 6, pp. 1840–1849, 2022, doi: 10.30865/jurikom.v9i6.5145.
- [10] E. Turban, J. E. Aronson, and T. P. Liang, *Decision Support Systems and Intelligent System*. 2005.
- [11] G. A. Pradyana and K. Agustini, *Basic Concepts of Data Mining*, Issue 1. 2021.
- [12] R. Tahir et al., "Business Inteligent (Introduction to Business Intelligence in Business)," 2023, pp. 48–61.
- [13] S. I. Nurhafida and F. Sembiring, "Sentiment Analysis of Online Novel Applications on Google Play Store Using Support Vector Machine (SVM) Algorithm," 2022.
- [14] Chambers and I. K. Sari, Eds., *Data Mining Textbook*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.
- [15] Amna et al., *Data Mining*. Padang: PT Global Ejecutivo Teknologi, 2023.
- [16] Indonesia Law No. 40 of 2004 concerning the National Social Security System, vol. 40, no. 150. Jakarta: Government of the Republic of Indonesia, 2004. [Online]. Available: <https://peraturan.bpk.go.id/Details/40787/uu-no-40->
- [17] P. Silitonga and I. S. Morina, "Clustering of Disease Spread Patterns Based on Patient Age Using K-Means Clustering," *TIMES Journal*, vol. 6, no. 2, 2018, doi: 10.51351/jtm.6.2.2017584.
- [18] A. Z. Syahputri, F. Della Fallenia, and R. Syafitri, "Quantitative Research Thinking Framework," *Tarbiyah: Journal of Education and Teaching*, 2023, [Online]. Available: <https://jurnal.diklinko.id/index.php/tarbiyah/>
- [19] Suggestion, *Qualitative, Quantitative, and R&D Research Methods*. Alphabet, 2022.
- [20] Moh. Nazir, *Research Methods*, Print 11. Bogor: Ghalia Indonesia, 2017.
- [21] S. Arikunto, *Research Procedure: A Practical Approach*, Ketu Edition. Jakarta: Rineka Cipta, 2019.
- [22] L. R. Gay, G. E. Mills, and P. W. Airasian, *Educational Research: Competencies for Analysis and Applications*, 10th ed. Upper Saddle River, NJ: Pearson, 2012.
- [23] I. Adiputra Mahayasa Nyoman, "Clustering of Dengue Disease at Dharma Kerti Hospital Using the K-Means Algorithm," *INSERT: Information System and Emerging Technology Journal*, vol. 2, no. 2, pp. 99–105, 2021.
- [24] S. Nurhayati and A. Fauzi, "Preparation of Research Roadmap as a Direction for Higher Education Research Development," *Scientific Journal of Education and Research*, vol. 5, no. 2, pp. 45–53, 2021.