PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.9032

CLUSTERING OF INPATIENTS USING THE K-MEANS METHOD AT RSIA SETYA BHAKTI

Yuliana Putri Ananda*; Al Fajri; Azka Jaisy Kahfi; Indah Suryani

Information Systems Study Program
Universitas Bina Sarana Informatika, Jakarta, Indonesia
www.bsi.ac.id
jrialfajri10@gmail.com, azkajaisy12@gmail.com, yulianaputri805@gmail.com, indah.ihy@bsi.ac.id
(*) Corresponding author



The creation is distributed under the Creative Commons AttributionNonCommercial 4.0 International License.

Abstract—*Inpatient data is a valuable source of information for identifying patterns in patient characteristics within a hospital. However, such data have not been fully utilized to analyze patient characteristics. This study aims to cluster inpatient data from RSIA Setya Bhakti based on age, length of stay, patient category, and disease group using the K-Means algorithm. The research stages included problem identification, data collection, data cleaning, preprocessing, data transformation, normalization, determination of the optimal number of clusters using the Elbow method, implementation of the K-Means algorithm using RapidMiner, and interpretation of the clustering results. The dataset consisted of 923 inpatient medical records from January to April 2026. Based on the Elbow method, K = 3 was selected as the optimal number of clusters, resulting in three distinct patient groups. Cluster 0 consisted of 443 patients (47.99%), predominantly pediatric patients, with an average age of 6.84 years, an average length of stay of 1.70 days, and a predominance of infectious diseases. Cluster 1 consisted of 152 patients (16.47%), predominantly patients in the maternal category, with an average age of 44.18 years, an average length of stay of 1.72 days, and more diverse disease characteristics. Cluster 2 consisted of 328 patients (35.54%), predominantly patients in the maternal category with pregnancy-related cases, with an average age of 31.17 years and an average length of stay of 1.71 days. The contribution of this study is the identification of distinct inpatient characteristic patterns through clustering, providing descriptive information that can support the hospital in understanding the distribution and characteristics of its inpatient population.*

Keywords: *clustering, data mining, disease groups, inpatients, KMeans.*

Abstrak—*Data pasien rawat inap merupakan sumber informasi yang dapat digunakan untuk mengidentifikasi pola karakteristik pasien di rumah sakit. Namun, data tersebut belum dimanfaatkan secara optimal untuk menganalisis karakteristik pasien. Penelitian ini bertujuan mengelompokkan data pasien rawat inap RSIA Setya Bhakti berdasarkan atribut umur, lama rawat inap, kategori pasien, dan kelompok penyakit menggunakan algoritma K-Means. Tahapan penelitian meliputi identifikasi masalah, pengumpulan data, pembersihan data, prapemrosesan, transformasi data, normalisasi, penentuan jumlah cluster optimal menggunakan metode Elbow, penerapan algoritma K-Means menggunakan RapidMiner, serta interpretasi hasil clustering. Dataset yang digunakan terdiri dari 923 rekam medis pasien rawat inap periode Januari hingga April 2026. Berdasarkan metode Elbow, K = 3 dipilih sebagai jumlah cluster optimal sehingga menghasilkan tiga kelompok pasien yang berbeda. Cluster 0 terdiri dari 443 pasien (47,99%) yang didominasi pasien kategori anak, dengan rata-rata usia 6,84 tahun, rata-rata lama rawat inap 1,70 hari, dan didominasi penyakit infeksi. Cluster 1 terdiri dari 152 pasien (16,47%) yang didominasi pasien kategori ibu, dengan rata-rata usia 44,18 tahun, rata-rata lama rawat inap 1,72 hari, serta memiliki karakteristik penyakit yang lebih beragam. Cluster 2 terdiri dari 328 pasien (35,54%) yang didominasi pasien kategori ibu dengan kasus terkait kehamilan, dengan rata-rata usia 31,17 tahun dan rata-rata lama rawat inap 1,71 hari. Kontribusi penelitian ini adalah mengidentifikasi pola karakteristik pasien rawat inap yang berbeda melalui clustering sehingga memberikan informasi deskriptif yang dapat membantu rumah sakit dalam memahami distribusi dan karakteristik populasi pasien rawat inap.*

Kata Kunci: *clustering; data mining; kelompok penyakit; pasien rawat inap, Kmeans.*

INTRODUCTION

Hospitals are healthcare facilities that provide comprehensive medical services, including emergency, outpatient, inpatient, referral, medical record, and medical support services, as well as education, training, and research activities (Yudano et al., 2022). Based on the services provided, hospitals can be categorized into general hospitals and specialty hospitals. Maternal and Child Hospitals (RSIA) are specialty healthcare facilities that focus primarily on maternal and pediatric healthcare services (Regulation of the Minister of Health of the Republic of Indonesia Number 3 of 2020). The number of hospitals in Indonesia has continued to grow, with Maternal and Child Hospitals accounting for a substantial proportion of specialty hospitals (Oktavia & Prayoga, 2023). This development highlights the importance of effectively utilizing hospital data to better understand the characteristics of the populations served.

Medical records contain valuable information regarding patient characteristics, examinations, treatments, procedures, and other healthcare services received by patients (Boris et al., 2024). In addition to supporting administrative activities, medical record data can be further utilized through data mining techniques such as K-Means clustering to identify patterns and group hospital patients based on similar characteristics (Ardiansyah et al., 2025). Effective utilization of such data can provide additional information for healthcare management and service evaluation (Widayanti et al., 2023). However, inpatient medical record data at RSIA Setya Bhakti in Depok have primarily been used for routine administrative and reporting purposes, while patterns within patient characteristics have not been systematically analyzed.

Clustering is a data mining technique that can be used to identify groups of observations with similar characteristics. Among clustering techniques, the K-Means algorithm is widely used because it partitions data into clusters based on the similarity of data points to their respective cluster centroids (Prastiwi & Mulyati, 2023); (Commodus et al., 2022). Previous studies have demonstrated the application of clustering and K-Means to healthcare and patient data (Idham et al., 2024). Similarly, (Sjamsuddin et al., 2025) applied the K-Means method to cluster inpatient BPJS patient data at RS Islam Assyifa Sukabumi. However, the characteristics and attributes used in clustering differ depending on the research context and the information required by healthcare facilities. However, the characteristics and attributes used in clustering differ depending on the research context and the information required by healthcare facilities. Existing applications do not specifically address the combination of age, length of stay, patient category, and disease group for characterizing the inpatient population of a Maternal and Child Hospital, particularly at RSIA Setya Bhakti.

Therefore, a research gap remains in the utilization of inpatient medical record data to characterize patient groups by simultaneously considering demographic, hospitalization, patient-category, and disease-group attributes in the specific context of a Maternal and Child Hospital. In this study, the patient category is derived during preprocessing based on the attending physician: patients treated by pediatricians (Sp.A) are categorized as pediatric patients, whereas patients treated by other relevant physicians are categorized into the maternal patient category. This derived attribute is combined with age, length of stay, and disease group to provide a more comprehensive representation of inpatient characteristics.

This study aims to apply the K-Means algorithm to cluster inpatient data from RSIA Setya Bhakti based on age, length of stay, patient category, and disease group. The novelty of this study lies in the integration of these four attributes, particularly the derived patient-category attribute based on the attending physician, to characterize inpatient groups within the specific context of a Maternal and Child Hospital. Unlike approaches that analyze patient characteristics using general demographic or clinical variables alone, this study incorporates patient category and disease group alongside age and length of stay to identify distinct patterns within the hospital's inpatient population. The resulting clusters are intended to provide descriptive information regarding inpatient characteristics and to improve the utilization of medical record data beyond routine administrative reporting.

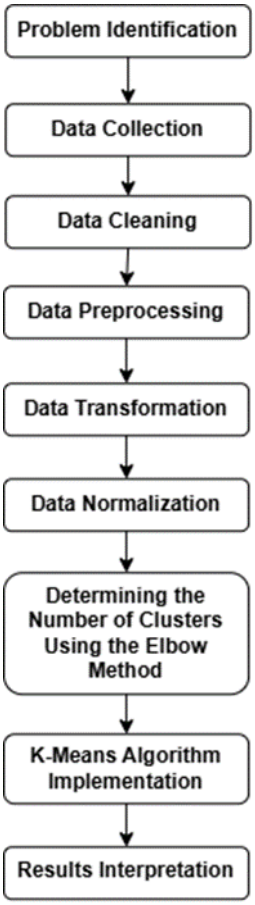
RESEARCH METHODOLOGY

This study employs a descriptive quantitative approach based on the KMeans Clustering algorithm. The data used in this research were obtained directly from RSIA Setya Bhakti in Depok.

The records comprise patient medical records and administrative data from inpatient stays during the period from January to April 2026. The initial dataset contained 2,539 records. Data selection and preprocessing were performed before the clustering analysis to ensure that the records contained complete and relevant information for the selected clustering attributes. The inclusion criteria consisted of inpatient records within the

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.9032

study period that contained valid information for age, length of stay, attending physician, and diagnosis required to determine the patient category and disease group. Records with missing or incomplete values in the required attributes, records that could not be categorized into the defined patient or disease groups, and records that did not meet the study criteria were excluded. Duplicate records were also identified and removed to prevent the same patient record from being represented more than once in the analysis. After the data cleaning and preprocessing stages, 923 records that met the inclusion criteria and contained complete information for the clustering variables were retained for further analysis. Four characteristics were subsequently used in the clustering process: age, length of stay, patient category, and disease group. The overall stages of the research process are illustrated in Figure 1.



Source: (Research Result, 2026)

Figure 1. Research Framework

A patient category attribute was introduced during the preprocessing stage by dividing the inpatient records into two groups: pediatric and maternal patients. This categorization was based on the patient groups represented in the dataset used in this study. Patients treated by pediatricians (Sp.A) were categorized as pediatric patients, while the remaining records in the analyzed dataset represented maternal patients and were categorized accordingly. Therefore, the "maternal" label does not imply that all patients treated by non-pediatricians are generally maternal patients; rather, it reflects the specific composition and scope of the RSIA Setya Bhakti inpatient dataset used in this study. This categorization was applied to distinguish the characteristics of pediatric and maternal patients in the clustering analysis included age, length of stay, category, and disease group. Subsequently, data processing and cluster formation were performed using RapidMiner.

RapidMiner is a robust software platform for machine learning and data science. The platform provides a wide range of tools for modeling, data preparation, evaluation, and deployment. RapidMiner's userfriendly interface enables users even those without programming skills to create and test various models (Nahjan et al., 2023)

RapidMiner is one of the programs used for data processing and analysis. It offers solutions for text mining, data mining, and predictive analysis (Idham et al., 2024)

RapidMiner/Altair AI Studio and Microsoft Excel were used to process the data in this study. Raw data were cleaned, their characteristics analyzed, and data categories adjusted using Microsoft Excel. The KMeans method was then employed for attribute selection, data normalization, and clustering, while the clustering results were evaluated using RapidMiner/Altair AI Studio.

Clustering is a data mining technique used to group data based on similarities in their characteristics. In healthcare data analysis, K-Means clustering can be applied to medical record data to identify groups of patients with similar characteristics (Laksono et al., 2024)

One of the advantages of the KMeans method which falls under the category of partitioning clustering is its applicability to problems involving largescale datasets (Prakoso et al., 2023) As KMeans is an unsupervised learning technique, labels or targets are not required during the clustering process. The following outlines the procedure for using the KMeans approach for clustering (Dewi et al., 2022):

1. Perform data transformation by converting categorical attributes into numerical ones. In this study, the patient category and disease group attributes were transformed using the Nominal to Numerical operator in RapidMiner/Altair AI Studio with the unique integer method. Each nominal disease category was assigned a unique numerical code from 0 to 8. The numerical codes were used as

computational representations of the disease categories and were not intended to indicate disease severity, ranking, or an ordinal clinical relationship. This encoding was retained to preserve consistency and reproducibility with the preprocessing procedure implemented in this study. However, because K-Means uses Euclidean distance, sequential integer encoding may introduce artificial numerical distances between nominal disease categories. Therefore, numerical proximity between disease codes was not interpreted as clinical similarity, and this limitation was considered when interpreting the clustering results.

2. Perform data normalization using the MinMax Normalization method so that all attributes share the same value range, ensuring no single attribute dominates the distance calculation. The normalization formula is as follows:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where:
X' = normalized value
X = original value
X_min = minimum value of the attribute
X_max = maximum value of the attribute

3. Determine the value of k, representing the number of clusters to be created.
4. There are other ways to initialize these k cluster centers, but selecting them randomly from the available data is the most common approach.
5. Use the Euclidean distance algorithm to determine the distance of each data point from each cluster center until the shortest distance between each data point and a cluster center is found. The Euclidean distance formula is as follows:

$$De = \sqrt{x_i - s_i)^2 (y_{i-} t_i)^2} \quad (2)$$

where:
De is the Euclidean Distance,
i is the number of objects,
(x, y) represents the object coordinates, and
(s, t) represents the centroid coordinates.

6. Sort each data point based on its proximity to the cluster center (shortest distance).
7. Update the cluster center values. Using the following formula, the new cluster center values are obtained from the average of the respective cluster:

$$v_{ij} = \frac{1}{N_i} \sum_{K=0}^{N_i} X_{kj} \quad (3)$$

where:
v_{ij} is the centroid (mean) of the ith cluster for the jth variable
N_i is the number of data points belonging to the ith cluster
i, k are cluster indices
j is the variable index
X_{kj} is the value of the kth data point within that cluster for the jth variable

8. Repeat steps b through e until the members of each cluster remain unchanged.

In this study, K-Means clustering was implemented using RapidMiner Studio version 10.2.000 with K = 3. The "determine good start values" option was enabled for centroid initialization, with a maximum of 10 runs. The measure type was set to Bregman Divergences using Squared Euclidean Distance, and the maximum number of optimization steps was set to 100. The local random seed option was not enabled; therefore, no fixed local random seed was specified.

RESULTS AND DISCUSSION

Medical record data were processed manually using Microsoft Excel. The secondary data used in this study were sourced from a medical record database containing 2,539 records 923 of which were analyzed and were presented in tabular form using Microsoft Excel (Dilawati et al., 2024)

In this study, the initial dataset comprises several attributes such as age, length of hospital stay, patient category, and disease group which serve as factors in the clustering process. Prior to clustering, categorical attributes (specifically patient category and disease group) are transformed into numerical form during the preprocessing stage to enable processing by the KMeans algorithm. A sample of the medical record dataset used in the clustering process is presented in Table 1.

Table 1. Medical Record Dataset

NO_RM	Age	Length_of_Stay	Category	Disease_Group
1001	29	0	Mother	Pregnancy
1002	6	1	Child	Digestive Diseases
1003	32	5	Mother	Respiratory Diseases
1004	13	2	Child	Infectious Diseases
1005	23	3	Mother	Digestive Diseases
1006	2	2	Child	Neurological Diseases
1007	72	1	Mother	Neurological Diseases

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.9032

NO_RM	Age	Length_of_Stay	Category	Disease_Group
1008	34	0	Mother	Neurological Diseases
1009	26	0	Mother	Pregnancy
1010	2	0	Child	Digestive Diseases
1011	2	2	Child	Digestive Diseases
1012	4	2	Child	Respiratory Diseases
1013	60	2	Mother	Metabolic Diseases
1014	2	4	Child	Digestive Diseases
1015	17	1	Mother	Respiratory Diseases
1016	61	0	Mother	Digestive Diseases
1017	63	0	Mother	Digestive Diseases
1018	1	2	Child	Infectious Diseases
1019	4	0	Child	Infectious Diseases
1020	3	0	Child	Infectious Diseases
1021	26	3	Mother	Pregnancy

Source: (Research Result, 2026)

Subsequently, the KMeans method groups the data based on characteristic similarities across the four categories. The dominant traits within each group derived from a combination of age, length of hospital stay, patient category, and disease group are then determined through an analysis of the resulting clusters.

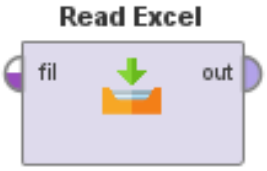
Problem Identification

Initial inpatient data includes various administrative characteristics, such as medical record number, patient name, address, telephone number, payer, room class, admission date, discharge date, attending physician, diagnosis, and discharge notes. The data utilized in this study comprises four attributes: age, length of stay, patient category, and disease group. The patient category is an additional attribute created during the preprocessing stage based on the attending physician. Patients treated by pediatricians (Sp.A) are categorized as pediatric patients, whereas those treated by physicians other than pediatricians are categorized as maternal patients. Subsequently, the patient category attribute along with the attributes of age, length of stay, and disease group serves as a variable in the clustering process using the KMeans algorithm.

Stages Of Data Mining

1. Data Selection

At this stage, the researcher uses the Read Excel operator in RapidMiner to import the dataset. The dataset was imported into RapidMiner using the Read Excel operator, as shown in Figure 2.

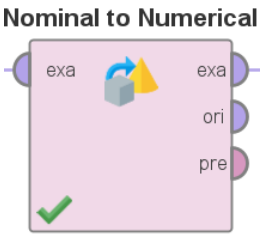


Source: (Research Result, 2026)

Figure 2. Read Excel operator

2. Data Transformation

The process of modifying data using the Nominal to Numerical operator in RapidMiner employing the unique integer method is used to convert nominal data into numerical data for the patient category and disease group attributes. The transformation of nominal attributes into numerical values was performed using the Nominal to Numerical operator, as shown in Figure 3.



Source: (Research Result, 2026)

Figure 3. Nominal to Numerical Operator

The numerical encoding of the disease groups and patient categories is presented in Tables 2 and 3, respectively.

Table 2. Results of Disease Data Transformation

Disease	Initialization
Pregnancy	0
Digestive Diseases	1
Respiratory Diseases	2
Infectious Diseases	3
Neurological Diseases	4
Metabolic Diseases	5
Heart Diseases	6
Tumors	7
Eye Diseases	8

Source: (Research Result, 2026)

Table 3. Results of Categorical Data Transformation

Category	Initialization
Mother	0
Child	1

Source: (Research Result, 2026)

After the categorical attributes were converted into numerical values, the resulting transformed dataset is presented in Table 4.

Table 4. Transformed dataset

NO_RM	Age	Length_of_Stay	Category	Disease_Group
1001	29	0	0	0
1002	6	1	1	1
1003	32	5	0	2
1004	13	2	1	3
1005	23	3	0	1
1006	2	2	1	4
1007	72	1	0	4
1008	34	0	0	4
1009	26	0	0	0
1010	2	0	1	1
1011	2	2	1	1
1012	4	2	1	2
1013	60	2	0	5
1014	2	4	1	1
1015	17	1	0	2
1016	61	0	0	1
1017	63	0	0	1
1018	1	2	1	3
1019	4	0	1	3
1020	3	0	1	3
1021	26	3	0	0

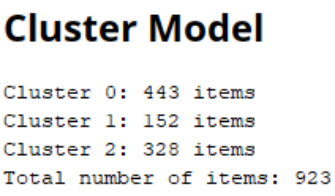
Source: (Research Result, 2026)

3. Perform Data Normalization

Data normalization was performed using the Normalize operator in RapidMiner/Altair AI Studio with the range transformation method. The minimum and maximum values were set to 0.0 and 1.0, respectively, so that all numerical attributes were transformed into the range of 0 to 1. This Min-Max normalization was applied to place the attributes on a comparable scale and prevent attributes with larger numerical ranges from dominating the Euclidean distance calculation in the K-Means clustering process.

4. Calculating The Centroid

In the clustering stage, the number of clusters (K) is determined, and the KMeans algorithm initializes centroids as cluster centers. Subsequently, each data point is assigned to a cluster based on its nearest distance to a centroid, calculated using the Euclidean distance metric. The process continues, with centroid positions updated based on the mean values of the cluster members, until convergence is reached. After preprocessing and normalization, the K-Means clustering process was performed in RapidMiner, and the resulting cluster model is shown in Figure 4.



Source: (Research Result, 2026)

Figure 4. KMeans clustering results.

Table 5. Cluster Centroids on the Normalized Scale (0–1)

Attribute	Cluster 0	Cluster 1	Cluster 2
Category	1.000000	0.000000	0.000000
Disease_Group	0.289414	0.477796	0.076600
Age	0.072982	0.469974	0.331571
Length_of_Stay	0.121299	0.122650	0.122168

Source: (Research Result, 2026)

The centroid values presented in Table 5 are reported on the normalized scale (0–1) resulting from the Min-Max normalization applied before K-Means clustering. Therefore, the centroid values for Age and Length of Stay represent normalized values rather than their original units of years and days. For the categorical attributes, Category and Disease Group, the centroid values should be interpreted cautiously because these attributes were transformed using unique integer encoding before normalization. Consequently, their centroid values represent numerical averages of the encoded values within each cluster and should not be interpreted as actual clinical categories, ordinal relationships, or measures of clinical similarity.

Cluster Results

The characteristics and distribution of the three clusters resulting from the K-Means analysis are summarized in Table 6.

Table 6. Clustering Results

Description	Cluster 0	Cluster 1	Cluster 2
Data volume	443	152	328
Age range	0 18 years	18 – 94 years	15 – 78 years
Patient category	Mother= 0 Child = 443	mother = 152 Child = 0	mother = 328 Child = 0
Type of disease	Infectious Diseases = 254 Metabolic Diseases = 2 Digestive Diseases = 129 Respiratory Diseases = 54 Neurological Diseases = 4	Infectious Diseases = 58 Heart Disease = 5 Eye Disease = 1 Metabolic Diseases = 35 Respiratory Diseases = 15 Neurological Diseases = 34 Tumors = 4	Pregnancy = 171 Digestive Diseases = 113 Respiratory Diseases = 44
Length of stay	0 – 9 day	0 – 8 day	0 – 14 day

Source: (Research Result, 2026)

The clustering results demonstrate distinct patterns in the inpatient population of RSIA Setya Bhakti. Cluster 0 is characterized by pediatric patients with a predominance of infectious diseases,

PILAR Nusa Mandiri: Journal of Computing and Information System
Vol. 22, No. 2 September 2026 | DOI: 10.33480/pilar.v22i2.9032

while Clusters 1 and 2 primarily represent maternal patients with different age and disease characteristics. Cluster 1 shows greater disease diversity among patients with a higher average age, whereas Cluster 2 is characterized by patients of reproductive age with a predominance of pregnancy-related cases. These patterns indicate that age, patient category, and disease group contribute to differentiating the characteristics of the inpatient population, while the average length of stay is relatively similar across the three clusters.

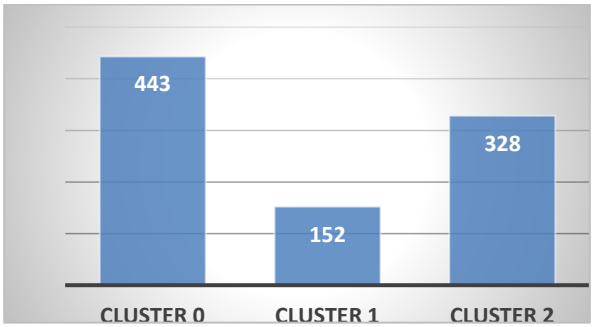
These findings are consistent with previous studies demonstrating the usefulness of K-Means for identifying patterns in medical record and patient data. (Dilawati et al., 2024) applied K-Means clustering to medical record data at Widodo Hospital Ngawi, while (Herdiaman et al., 2024) used K-Means to cluster patient data at RSUD Ciamis. Similarly, (Yudano et al., 2022) demonstrated the use of medical record data to identify disease clusters through data mining. These studies support the use of clustering for extracting patterns from medical record datasets that may not be readily apparent from routine administrative reporting. The present study extends this application to the specific context of a Maternal and Child Hospital by considering age, length of stay, patient category, and disease group simultaneously.

The observed cluster patterns can be understood in relation to the characteristics of the patient population represented in the dataset. The separation of pediatric patients in Cluster 0 is associated with the pediatric category and their younger age distribution, accompanied by a high proportion of infectious diseases. In contrast, the maternal patients are separated into two clusters with different age and disease profiles. Cluster 1 contains patients with a higher average age and a more heterogeneous distribution of diseases, whereas Cluster 2 contains younger maternal patients and is dominated by pregnancy-related cases. However, these patterns should be interpreted as descriptive characteristics of the analyzed dataset rather than as evidence of causal relationships between age, patient category, and disease occurrence.

From a hospital data management perspective, the clustering results demonstrate how routinely collected inpatient medical record data can be transformed into descriptive information about patient population patterns. Instead of being used solely for administrative reporting, the data can be summarized into patient groups that allow the hospital to better understand the distribution of pediatric and maternal patients and their associated disease characteristics. Such information may support internal data monitoring, evaluation of patient distributions, and identification of frequently occurring disease groups.

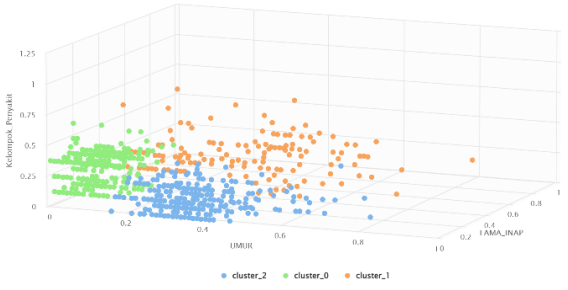
From a service-planning perspective, the identified clusters can provide supplementary descriptive information for understanding differences in service needs across patient groups. For example, the predominance of infectious diseases among pediatric patients and pregnancy-related cases among younger maternal patients highlights different patterns within the hospital's inpatient population. Nevertheless, the clustering results should be regarded as supporting descriptive information rather than direct evidence for determining specific clinical interventions or service programs. Further evaluation incorporating additional clinical and operational variables would be required before the clusters could be used to guide specific hospital service-planning decisions.

Interpretation Of Results



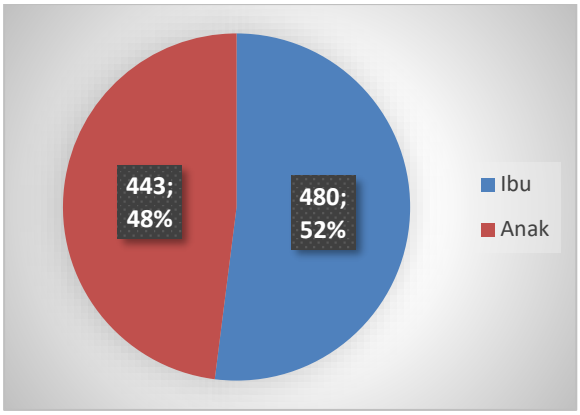
Source: (Research Result, 2026)
Figure 5. Visualization of clustering results.

Based on Figure 5, the clustering results are visualized using a bar chart to illustrate the differences in the number of patients across the three clusters. The K-Means process with K = 3 resulted in Cluster 0 comprising 443 patients (47.99%), Cluster 1 comprising 152 patients (16.47%), and Cluster 2 comprising 328 patients (35.54%). The unequal distribution indicates that the inpatient records are characterized by patient groups of different sizes rather than being evenly distributed across the three clusters.



Source: (Research Result, 2026)
Figure 6. Scatter plot

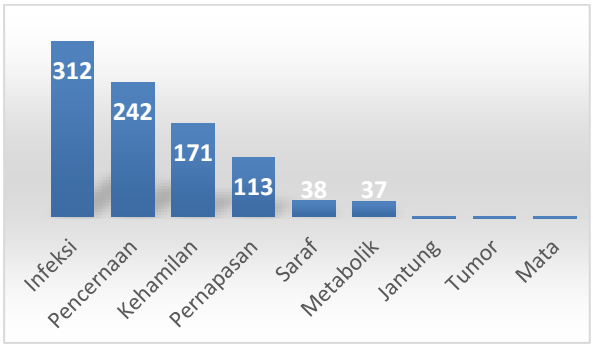
Figure 6 visualizes the patient data grouped into three clusters using the K-Means algorithm. Each point represents a patient record, while the different colors represent the resulting clusters. The visualization illustrates the separation of patient records according to similarities in the attributes used in the clustering process. The grouping is determined by the proximity of each data point to its cluster centroid, resulting in patients with more similar characteristics being assigned to the same cluster.



Source: (Research Result, 2026)

Figure 7. Results of patient category grouping.

Figure 7 presents the distribution of patient categories in the analyzed dataset. The maternal category accounts for a larger proportion than the pediatric category. This distribution reflects the composition of inpatient records used in the study and provides additional context for interpreting the characteristics of the clusters, particularly the predominance of maternal patients in Clusters 1 and 2.



Source: (Research Result, 2026)

Figure 8. Results of disease type grouping.

Figure 8 presents the distribution of disease groups recorded in the inpatient dataset. The disease groups include infectious, digestive, pregnancy-related, respiratory, neurological, metabolic, heart, tumor, and eye-related conditions. Differences in the distribution of these disease groups contribute to the characteristics of the resulting clusters, particularly the predominance of infectious diseases among pediatric patients and pregnancy-related cases among maternal patients of reproductive age.

The clustering results reveal distinct patterns among the inpatient population of RSIA Setya Bhakti. Cluster 0 is predominantly composed of pediatric patients with a relatively young average age and a high proportion of infectious diseases. Cluster 1 represents maternal patients with a higher average age and a more heterogeneous distribution of diseases, particularly infectious, metabolic, and neurological diseases. In contrast, Cluster 2 mainly consists of maternal patients of reproductive age and is characterized by a predominance of pregnancy-related cases. Although the clusters differ in terms of age, patient category, and disease characteristics, the average length of stay is relatively similar across the three clusters. This suggests that length of stay provides less differentiation among the identified patient groups than the other attributes used in the analysis.

These findings are consistent with previous studies demonstrating that K-Means clustering can be used to identify patterns within medical record and patient datasets. Dilawati, Widianto, and Kuswiadji (2024) applied K-Means clustering to medical record data at Widodo Hospital Ngawi, while Herdiaman (2024) applied the method to patient data at RSUD Ciamis. Yudano, Indahyanti, and Nisak (2022) also demonstrated the use of medical record data in disease clustering. Similar to these studies, the present findings demonstrate the potential of clustering to transform routinely collected patient data into identifiable patient groups. However, the present study focuses specifically on the context of a Maternal and Child Hospital and incorporates age, length of stay, patient category, and disease group to characterize its inpatient population.

The observed patterns may be related to the specific characteristics of the hospital population represented in the dataset. The formation of a predominantly pediatric cluster is associated with the younger age distribution and the concentration of infectious diseases among pediatric records. Meanwhile, maternal patients are divided into two groups with different age and disease profiles. The higher average age and greater disease diversity in Cluster 1 distinguish it from Cluster 2, which consists primarily of younger maternal patients and has a high proportion of pregnancy-related cases. These interpretations describe associations observed within the dataset and should not be considered evidence of causal relationships between patient age and disease occurrence.

From a hospital data management perspective, these results demonstrate that routinely collected inpatient medical records can be utilized beyond administrative reporting. Clustering can summarize a relatively large number of patient records into groups with identifiable characteristics, making it easier to understand the distribution of pediatric and maternal patients and the disease patterns associated with these groups. The resulting information may support internal data monitoring, evaluation of patient distributions, and identification of frequently occurring disease groups.

From a service-planning perspective, the identified patterns may provide supplementary information for understanding differences among inpatient groups. For example, the predominance of infectious diseases among pediatric patients and pregnancy-related cases among maternal patients of reproductive age may help the hospital identify patient groups and disease categories that frequently appear in inpatient records. However, these clustering results should be regarded as descriptive supporting information rather than direct evidence for determining specific clinical interventions or resource allocations. Additional clinical and operational variables, as well as further validation, would be required before the clustering results could be used for specific hospital service-planning decisions.

CONCLUSION

This study identified three distinct inpatient clusters at RSIA Setya Bhakti based on age, length of stay, patient category, and disease group. Cluster 0 comprised 443 patients (47.99%), predominantly pediatric patients with infectious diseases; Cluster 1 comprised 152 patients (16.47%), characterized by patients with a higher average age and more diverse disease conditions, particularly infectious, metabolic, and neurological diseases; and Cluster 2 comprised 328 patients (35.54%), predominantly maternal patients of reproductive age with pregnancy-related cases. These findings provide descriptive information that can support the hospital in understanding inpatient characteristics and patient distribution for internal data monitoring and service planning. However, this study is limited to 923 inpatient records from January to April 2026 and four clustering attributes, which may limit the generalizability of the identified patterns. Future studies should use a longer observation period, larger datasets, and additional clinical and operational attributes, as well as evaluate alternative clustering methods to assess the robustness of the resulting patient groups.

REFERENCE

Ardiansyah, A. A., Khaeroshi, E. A., & Wicaksono, R. R. (2025). Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di RSUD Muhammadiyah Bantul. *JUTIS*, *13*(2), 141–150. https://www.ejournal.unis.ac.id/index.php/jutis/article/view/6776

Boris, J., Ginting, A., Gulo, V. E., Zebua, R., Nainggolan, M., & Tarigan, M. W. (2024). Peran Rekam Medis sebagai Penyedia Informasi Kesehatan pada Siswa- Siswi SMA Negeri 1 Parbuluan. *Inovasi Jurnal Pengabdian Masyarakat*, *2*(1), 51–58.

Commodus, P. E., Mandiri, U., Prestasi, B., & Penerbangan, P. (2022). Analisis Metode K-Nearest Neighbor Menggunakan Rapid Miner Untuk Sistem Rekomendasi Tempat Wisata Labuan Bajo. *Remik: Riset Dan E-Jurnal Manajemen Informatika Komputer*, *6*(3), 534–541.

Dewi, Y. N., Rianto, H., Budihartanti, C., & Fibriany, F. W. (2022). PENERAPAN METODE K-MEANS DALAM UJIAN NASIONAL. *Journal of Information System, Applied, Management, Accounting and Research*, *6*(1), 26–31. https://doi.org/10.52362/jisamar.v6i1.670

Dilawati, H., Widianto, H., & Kuswiadji, A. (2024). Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering Di Rumah Sakit Widodo Ngawi. *Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), 139–147.

Herdiaman, E. A., Sudiarjo, A., & Hikmatyar, M. (2024). KLASTERISASI PASIEN PADA RSUD CIAMIS MENGGUNAKAN METODE K-MEANS. *JITET (Jurnal Informatika Dan Teknik Elektro Terapan)*, *12*(3), 3538–3547.

Idham, Rosika, H., & Yuliadi. (2024). IMPLEMENTASI RAPIDMINER UNTUK CLESTERING DATA PENJUALAN PAKAIAN MENGGUNAKAN METODE K-MEANS. *Journal Education and Technology*, *5*(1), 221–231.

Laksono, W. B., Syahidin, Y., & Yunengsih, Y. (2024). Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *7*(2), 621–627. https://doi.org/10.32493/jtsi.v7i2.39354

Nahjan, M. R., Heryana, N., & Voutama, A. (2023). IMPLEMENTASI RAPIDMINER DENGAN METODE CLUSTERING K-MEANS UNTUK ANALISA PENJUALAN PADA TOKO OJ CELL. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 101–104.

Oktavia, S. N., & Prayoga, D. (2023). KUALITAS PELAYANAN TERHADAP LOYALITAS PASIEN

PADA RUMAH SAKIT IBU DAN ANAK : LITERATURE REVIEW. *JURNAL KESEHATAN TAMBUSAI*, *4*(3), 2199–2205.

Prakoso, B. H., Rachmawati, E., Putro, R. D. M., Vestine, V., & Suyoso, G. E. J. (2023). Klasterisasi Puskesmas dengan K-Means Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat. *Jurnal Buana Informatika*, *14*(April), 60–68.

Prastiwi, O., & Mulyati, S. (2023). PENERAPAN ALGORITME K-MEANS CLUSTERING DALAM PENGELOMPOKAN PENYAKIT PASIEN PADA UPTD PUSKESMAS WURYANTORO. *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, *2*(2), 577–586.

Sjamsuddin, I. N., Firmansyah, D. A., & Laferani, Y. (2025). Klasterisasi Pasien Rawat Inap BPJS pada RS Islam Assyifa Sukabumi Menggunakan Metode K-Means. *Jurnal Kridatama Sains Dan Teknologi*, *7*(01), 355–368. https://doi.org/10.53863/kst.v7i01.1617

Widayanti, E., Septiana, D. H., Irmaningsih, M., Putri, V. A., & Budi, S. C. (2023). Kesiapan Puskesmas Samigaluh I dalam Peralihan Rekam Medis Konvensional ke Rekam Medis Elektronik. *Jurnal Manajemen Informasi Kesehatan Indonesia*, *11*(2), 102–107.

Yudano, B. A., Indahyanti, U., & Nisak, U. K. (2022). Pemanfaatan Data Rekam Medis dalam Menentukan Cluster Penyakit Melalui Data Mining di Rs Aisyiyah Siti Fatimah Tulangan Sidoarjo. *Care Journal*, *2*(1), 14–23.