International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 4, Number 5, January 2026, *Page. 1547 - 1552*
Email : editorijhess@gmail.com

Disease Clusterization Based on Patient Age and Disease Type Using K-Means Clustering

Jalaluddin Mahally Hasibuan [1)*], Hendra Cipta [2)], Rini Halila Nasution [3)]
[1,2,3)] Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara

*Corresponding Author
Email : jalaluddinmahally9@gmail.com

Abstract

This study aims to classify disease types based on patient age using the K-Means Clustering method in order to identify disease distribution patterns at Malahayati Islamic Hospital, Medan. The data used in this research consists of medical record data of patients from October to December 2024, including variables such as age, type of disease, gender, and area of residence. The research stages include data cleaning, data transformation of age and disease attributes into numerical values, and clustering analysis using the K-Means algorithm implemented through RapidMiner software. The clustering results produced three main clusters, representing high, moderate, and low disease prevalence levels. Diseases with the highest prevalence cluster include pregnancy-related cases, pneumonia, acute respiratory infections (ISPA), chronic obstructive pulmonary disease (COPD), and gastroenteritis (GEA), which are predominantly found in adult and elderly age groups. The results indicate that patient age significantly influences disease distribution patterns. This study demonstrates that K-Means Clustering is effective in identifying age-based disease patterns and can serve as a decision-support tool for healthcare planning, resource allocation, and disease prevention strategies in hospital management.

Keywords: Data Mining, Disease Classification, Hospital Data, K-Means Clustering, Patient Age

INTRODUCTION

Health is a fundamental aspect of human life that directly affects individual quality of life and overall societal well-being. Disease patterns among patients vary significantly and are strongly influenced by age and type of illness. Each age group exhibits distinct health characteristics and risks, which require different clinical and preventive approaches. Therefore, systematic analysis of patient health data is essential to support evidence-based decision-making in healthcare services (Fajri & Purnamasari, 2022).

The increasing volume of medical record data stored in healthcare institutions has created challenges in extracting meaningful and actionable information. Without proper analytical methods, these large datasets remain underutilized. Data mining techniques provide an effective solution by enabling the discovery of hidden patterns and relationships within patient data, particularly in identifying disease distribution based on age groups and disease types. Such insights are crucial for improving healthcare planning and optimizing resource allocation (Lestari, 2022).

Clustering is one of the most widely applied data mining techniques in healthcare analytics due to its ability to group patients based on similarity without prior labelling. Disease clustering plays a significant role in public health by identifying trends and disease concentration across different age groups. Among various clustering algorithms, K-Means is favored for its computational efficiency, simplicity, and suitability for large-scale medical datasets, making it an effective tool for disease pattern analysis (Nabila et al., 2021; Okta, 2023).

The urgency of this research arises from the need for timely and data-driven health interventions, particularly in hospitals facing increasing patient loads and diverse disease profiles. Inaccurate or delayed identification of disease patterns may lead to suboptimal resource distribution, delayed preventive measures, and reduced quality of healthcare services. By applying K-Means clustering to patient data, healthcare providers can quickly identify high-risk age groups and dominant disease clusters, thereby supporting early intervention and improving service effectiveness (Rosida & Wijaya, 2023).

The novelty of this study lies in the integration of age-based disease clustering with real-world hospital data using a structured transformation process and cluster evaluation approach. Unlike previous studies that focus solely on age or disease type, this research simultaneously analyzes age, disease type, gender, and patient location to generate more comprehensive disease clusters. Additionally, the use of K-Means clustering combined with practical implementation tools provides new insights into disease distribution patterns that can directly support hospital-level decision-making and preventive health strategies (Sari et al., 2020).

RESEARCH METHODS

This research was conducted from October to December 2024. The research used a quantitative approach. Data were obtained from the Malahayati Islamic Hospital, Medan Petisah District, Medan City, North Sumatra, in 2011. In a study that utilized K-Means Clustering to group diseases based on patient age at Malahayati Islamic Hospital, several variables frequently used include: Patient age, type of disease such as chronic diseases such as diabetes, hypertension, coronary heart disease; infectious diseases such as tuberculosis, malaria, dengue fever; degenerative diseases such as arthritis, Alzheimer's, Parkinson's; gastrointestinal diseases such as gastritis, peptic ulcers; neurological diseases such as stroke, epilepsy; respiratory diseases such as asthma, chronic bronchitis; skin diseases such as dermatitis, psoriasis; gender, and location of residence.

K-Means Clustering Algorithm

K-Means Clustering is a method used in data mining which works by searching for and grouping data that has similar characteristics between one data and another. K-Means algorithm is relatively more scalable and efficient due to its relatively high precision regarding object size. Determining the centroid, number of clusters, and centroid distance is an important step in implementing K-Means Clustering (Rosida & Wijaya, 2023). The following are the steps in using the k-means method in clustering.

1. Determine k as the number of clusters to be formed.
2. Determine the initial k Centroids (cluster center points) randomly.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; \mathrm{i} = 1,2,3,...,n \qquad (1)$$

Where :

v : centroid of the cluster

x_i : object i

n : number of objects

3. Calculate the distance between each object and each centroid of each cluster. To calculate the distance between an object and the centroid, you can use the Euclidean distance.

$$\mathrm{d}_{ik} = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \qquad (2)$$

Where :

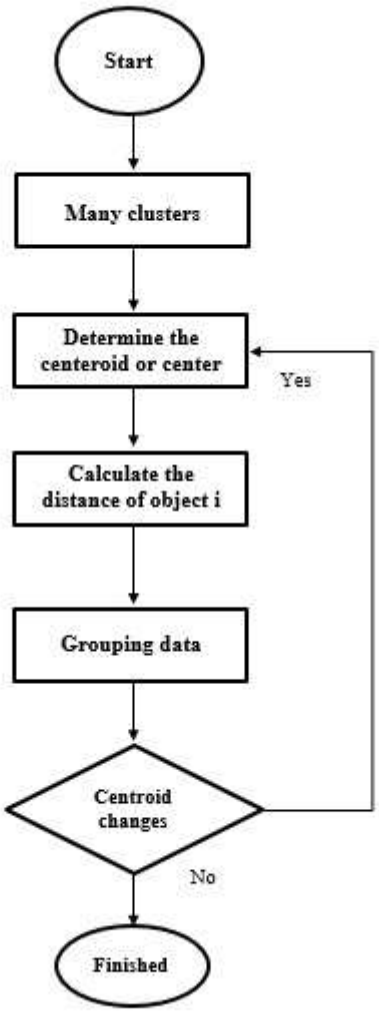
d_{ik} : distance between the i-th data and the cluster center point

m : number of attributes

x_{ij} : i-th data

c_{kj} : k-th class center data

The following is a research flowchart



RESULTS AND DISCUSSION

This study used a dataset of 100 patient visits at Malahayati Islamic Hospital from October to December 2024. The patient visit data was processed by removing inconsistent data, then transformed to convert the data from its initial form into a form suitable for grouping. Thus, the attributes selected for this study were month, name, age, disease type, gender, village, and medication, as shown in Table 1.

Tabel 1. Patient Visit Dataset

No	Month	Name	Gender	Age	Address	Type of disease
1	01.10.2024	Chairuddin	M	67	Medan	PPOK
2	01.10.2024	M Ali syahdana	M	68	Medan	HHD
3	01.10.2024	Jamilah	F	60	Medan	Anemia
4	01.10.2024	Martinah	F	58	Medan	Pneumonia
5	01.10.2024	Siti Fatimah	F	51	Deli Serdang	Anemia
⋮	⋮	⋮	⋮	⋮	⋮	⋮
97	05.10.2024	Annisa zahara hrp	F	21	Medan	ISPA
98	05.10.2024	Bayi Annisaa' adha savira	M	0	Medan	Kehamilan
99	05.10.2024	Nurul arrizka Putri	F	24	Medan	Pneumonia

100	05.10.2024	aqilla nazwa	F	19	Medan	Multisinusitis

The dataset of patient visits represents outpatient and inpatient records collected during the observation period. Each record contains patient demographic information and diagnosed disease data, which serve as the primary variables for clustering analysis. This dataset provides an overview of patient distribution across age groups and disease categories, forming the basis for identifying disease patterns using the K-Means algorithm. The completeness and diversity of the dataset allow for meaningful clustering results and support further analysis of age-related disease characteristics.

Then data transformation because according to the Ministry of Health of the Republic of Indonesia in 2009, the age categories are explained as toddlers, adolescents, adults, and elderly (elderly) divided into several categories including toddlerhood from 0–5 years old, childhood 6–11 years old, early adolescence 12–16 years old, late adolescence 17–25 years old, early adulthood 26–35 years old, late adulthood 36–45 years old, early elderly age 46–55 years old, late elderly age 56–65, and elderly age 65 years and above. Therefore, the transformation of the age range carried out in this study based on the age categories can be seen in the table 2.

Table 2. Data Transformation on Age Attribute

Variabel	Age range	e Grup	Frequency	Transformation value
Age	0-5 Tahun	Balita	199	1
	6-11 Tahun	Kanak-anak	81	2
	12-16 Tahun	Remaja Awal	49	3
	17-25 Tahun	Remaja Akhir	202	4
	26-35 Tahun	Dewasa Awal	321	5
	36-45 Tahun	Dewasa Akhir	162	6
	46-55 Tahun	Lansia Awal	223	7
	56-65 Tahun	Lansia Akhir	298	8
	>66 Tahun	Manula	289	9

After the age transformation, then the disease type transformation on the disease attribute is carried out an initialization process which is sorted based on the number of patients suffering from the disease at the Malahayati Islamic Hospital in October-December 2024. The Transformation Value is determined based on the frequency of the number of patients for each disease as can be seen in table 3.

Tabel 3. Data Transformation on Disease Attributes

No	Disease	Frequency	Transformation value
1	Kehamilan	218	1
2	Pneumonia	208	2
3	ISPA	183	3
4	PPOK	115	4
5	GEA	81	5
⋮	⋮	⋮	⋮
203	Traumatic amputasi jari III IV Tangan kiri	1	203
204	Tumor payudara kanan	1	204
205	Union clavicle with implant	1	205
206	Vertigo ec Multisinusitis	1	206

Next, in the Data Mining stage, data processing and pattern or information searching will be carried out using clustering techniques using the K-Means algorithm and data processing using RapidMiner. This involves a sample of 20 of the 1,221 data records in Table 4.

Tabel 4. Data Transformation

Patient	Age transformation	Gender transformation	Gender transformation	Regency transformation
1	9	2	4	1
2	9	2	23	1
3	8	1	6	1
4	8	1	2	1
5	7	1	6	2
⋮	⋮	⋮	⋮	⋮
97	4	1	3	1
98	1	2	1	1
99	4	1	2	1
100	4	1	40	1

After completing the data transformation, the next step is to classify the data using k-means. The classification results can be seen in the table 5.

Tabel 5. Classification results

Patient	C1	C2	C3	Closest Distance	Grup
1	5.6345579	91.145831	33.694304	5.6345579	1
2	15.300991	72.184226	35.460527	15.300991	1
3	3.5639487	89.115445	33.319113	3.5639487	1
4	6.789483	93.110485	34.065363	6.789483	1
5	2.7201848	89.056513	32.295826	2.7201848	1
⋮	⋮	⋮	⋮	⋮	⋮
97	5.4860805	92.08997	33.886666	5.4860805	1
98	8.5142485	94.183664	34.711346	8.5142485	1
99	6.4378503	93.089003	34.11565	6.4378503	1
100	31.877737	55.150363	44.476547	31.877737	1

The results of the study after data modeling were then analyzed, patient data was divided into 3 clusters where cluster 1 is a disease cluster with high sufferers, cluster 3 with moderate disease sufferers and cluster 2 with low disease sufferers. The highest disease groups are Pregnancy, Pneumonia, ISPA, COPD, GEA, Ca Mamae, and Viral Infection in October-December 2025. Researchers can identify medical record data from Malahayati Islamic Hospital as many as 1,824 patient data with a completion time of 0.06 seconds by the system.

CONCLUSIONS

This study concludes that the application of the K-Means Clustering method is effective in grouping diseases based on patient age using medical record data from Malahayati Islamic Hospital. The clustering process successfully formed three distinct disease groups representing high, moderate, and low levels of patient prevalence. The results indicate that certain diseases such as pregnancy-related cases, pneumonia, ISPA, COPD, and GEA dominate specific age groups, particularly adults

and the elderly. These findings demonstrate that age has a significant influence on disease distribution patterns. The clustering results can assist hospitals in identifying high-risk patient groups, improving healthcare service planning, optimizing medical resource allocation, and supporting early preventive measures. Therefore, K-Means Clustering can be utilized as a reliable data mining approach to support decision-making processes in healthcare management.

REFERENCES

Fajri, M. B., & Purnamasari, S. D. (2022). *Klasterisasi Pola Penyebaran Penyakit Pasien Berdasarkan Usia Pasien Menggunakan K-Means Clustering. Journal of Information Technology Ampera*, 3(3), 317–334. https://journalcomputing.org/index.php/journal-ita/index.

Lestari, S. (2022). *Penerapan Algoritma K-Means Untuk Pemetaan PenyebaranPenyakit Demam Berdarah (DBD) Pada Kabupaten/Kota Di Jawa Barat. Jurnal Pendidikan Dan Konseling*, 4, 1349–1358.

Nabila, Z., Rahman Isnain, A., & Abidin, Z. (2021). *Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma KMeans.* Jurnal Teknologi Dan Sistem Informasi *(JTSI)*, 2(2), 100. http://jim.teknokrat.ac.id/index.php/JTSI.

Okta Jaya Harmaja., Hadirat Halawa., (2023). *Implementasi Algoritma K-Means Clustering Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Pulo Brayan.* Jurnal Sains dan Teknologi, Vol. 5. No. 1

Rosida, W., & Wijaya, Y. A. (2023). *Klasterisasi Penyakit HIV/AIDS di Jawa Barat Menggunakan Algoritma K-Means Clustering*. Blend Sains Jurnal Teknik, 1(4), 306–315. https://doi.org/10.56211/blendsains.v1i4.235

Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). *Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer*. CESS (Journal of Computer Engineering, System and Science), 5(2), 192. https://doi.org/10.24114/cess.v5i2.18519