

Clustering the Severity Levels of Hypertensive Patients Based on Blood Pressure using the K-Means Method

Lia Anggraini^{1*}, Akim Manoor Hara Pardede², Kristina Annatasia Br Sitepu³

^{1,2,3} STMIK KAPUTAMA

liaanggraini030904@gmail.com^{1*}, Akimmhp@live.com², kannatasia88@gmail.com³

Abstract

Hypertension is a noncommunicable disease that is on the rise at the Anugrah Binjai Primary Care Clinic; however, the use of patient medical records has so far been limited to administrative purposes. This study aims to objectively and systematically classify the severity of hypertension based on data using data mining techniques. The algorithm applied is K-Means Clustering, utilizing the variables of gender, age, and systolic blood pressure from patients' medical records. The clustering process was processed and tested using the MATLAB application through Euclidean distance calculations until convergence was achieved. The results of the study using a three-cluster scenario produced a structured division of patient groups, namely Cluster 1 (moderate severity) with 151 data points, Cluster 2 (high severity) with 151 data points, and Cluster 3 (low severity) with 215 data points. An evaluation of clustering quality using cluster variance yielded values of 0.9359 for 3 clusters, 0.8412 for 4 clusters, and 0.7132 for 5 clusters. These results demonstrate that the K-Means method is capable of mapping the severity patterns of hypertensive patients to assist clinics in designing targeted educational programs and medical management.

Keywords: Data Mining, Hypertension, K-Means Clustering, MATLAB, Medical Records

1. Introduction

Hypertension is a major noncommunicable disease characterized by a persistent increase in arterial blood pressure. Without appropriate monitoring and management, this condition may increase the risk of cardiovascular disease, stroke, kidney failure, and other serious complications [1][2]. The clinical characteristics of hypertensive patients vary according to factors such as age, gender, and blood pressure level. This variation highlights the need for a systematic, objective, and data-driven approach to identifying groups of patients with similar severity characteristics[3].

At the Anugrah Binjai Primary Care Clinic, the number of patients with hypertension has shown an increasing tendency. However, the available medical records have primarily been used for administrative documentation rather than analytical purposes. Consequently, patient assessments are generally performed individually, without a structured overview of severity patterns across the entire patient population. This condition limits the clinic's ability to identify groups requiring different levels of monitoring, education, and preventive intervention[4]. Therefore, routine medical records need to be transformed into meaningful information that can support more targeted healthcare planning[5].

Data mining provides computational techniques for extracting patterns and relationships from large datasets that may not be identified through conventional administrative processing [6]. One technique commonly applied to medical data is clustering, which groups observations according to the similarity of their characteristics [7]. K-Means is a clustering algorithm that partitions data by assigning each observation to the nearest centroid through iterative distance calculations [8]. Its relatively simple computational process and interpretable cluster structure make K-Means suitable for exploring patterns within routinely collected patient records [9].

Previous studies have demonstrated the application of K-Means in grouping patients by disease distribution, hospital complaints, general disease risk, and heart disease characteristics. A hypertension-specific study also applied K-Means to age, systolic blood pressure, diastolic blood pressure, and body weight, but the resulting groups were limited to low- and high-severity categories [7]. Thus, the remaining research issue is not merely whether K-Means can process medical data, but whether it can produce an operationally interpretable three-level severity structure consisting of low, moderate, and high groups from routine primary-clinic records [10]. Studies that combine this three-level interpretation with an evaluation of several cluster configurations remain limited, particularly when the analysis is focused on gender, age, and systolic blood pressure [11].

Accordingly, this study applies the K-Means algorithm to 517 medical records of hypertensive patients collected at the Anugrah Binjai Primary Care Clinic from 2022 to 2025. The clustering process uses gender, age, and systolic blood pressure as analytical variables. A three-cluster configuration is interpreted into low, moderate, and high severity levels, while four- and five-cluster configurations are examined to compare cluster compactness using cluster variance. The study is expected to provide a structured mapping of hypertensive patient characteristics that can support patient monitoring and targeted health education. The resulting clusters are intended as data-driven decision-support information rather than as a substitute for clinical diagnosis [12].

2. Discussion

The K-Means clustering process was conducted to generate structured information regarding the severity characteristics of hypertensive patients at the Anugrah Binjai Primary Care Clinic. The analysis used gender, age, and systolic blood pressure as the principal variables. These variables were selected to identify groups of patients with similar characteristics and to provide an objective overview of hypertension severity patterns.

The research process consisted of data selection, data transformation, cluster-number determination, initial centroid assignment, distance calculation, data allocation, and centroid updating. The distance between each patient record and the cluster centroids was calculated using Euclidean Distance. Each record was then assigned to the nearest centroid, and the centroid values were updated iteratively until the cluster memberships became stable.

MATLAB R2014b was used to process the dataset and implement the K-Means algorithm. The clustering results were visualized in three-dimensional graphs, where the X-axis represented gender, the Y-axis represented age, and the Z-axis represented systolic blood pressure. Three-, four-, and five-cluster configurations were tested to compare the compactness and interpretability of the resulting patient groups.

2.1. Discussion of Input Data

The dataset was obtained from the medical records of hypertensive patients at the Anugrah Binjai Primary Care Clinic from 2022 to 2025. Although the original records contained several attributes, including patient name, gender, age, blood pressure, height, weight, and hypertension information, this study used only gender, age, and systolic blood pressure in the clustering process.

Before being processed, categorical and interval data were transformed into numerical values. This transformation was required because the K-Means algorithm calculates similarities between observations using numerical distances. The details of the input data are presented as follows:

File name : DataProses.xlsx
 Number of records : 517 patient records
 Observation period : 2022 - 2025
 Variabel : X = Age
 Y = Gender
 Z = Systolic blood pressure

The transformation values assigned to each variable are presented in Table 1.

Table 1: Transformation values of research variables

No	Variabel	Trasformasi	Nilai Transformasi
1	Jenis Kelamin (X)	Laki-Laki	1
		Perempuan	2
2	Usia (Y)	30-40 tahun	1
		41-50 tahun	2
		51-60 tahun	3
		61-70 tahun	4
		≥ 71 tahun	5
3	Tekanan Darah (Z)	140 – 150 mmHg	1
		151 – 160 mmHg	2
		161 – 170 mmHg	3
		≥ 171 mmHg	4

The transformation retained the ordinal characteristics of age and systolic blood pressure. A higher transformation value represented an older age group or a higher systolic blood pressure category. Gender was included as a supporting demographic variable rather than as the primary determinant of hypertension severity. The complete input configuration and transformation categories correspond to those applied to the 517 medical records in the study.

2.2. Implementation

Implementation was performed by importing the transformed dataset into MATLAB R2014b and organizing the three variables into a numerical matrix. The K-Means algorithm was then applied by determining the value of k , generating initial centroids, calculating the distance of each patient record from every centroid, and assigning each record to its nearest cluster.

For each iteration, the centroid of a cluster was recalculated from the average values of all records assigned to that cluster. The calculation was repeated until no further changes occurred in cluster membership. The final output consisted of cluster labels, centroid coordinates, the number of records in each cluster, and a three-dimensional visualization of the grouping results.

The primary analysis used three clusters because the resulting groups could be interpreted according to three hypertension severity levels: low, moderate, and high. Four- and five-cluster configurations were subsequently tested to examine whether increasing the number of clusters produced more compact and homogeneous groups.

1. Results of three clusters

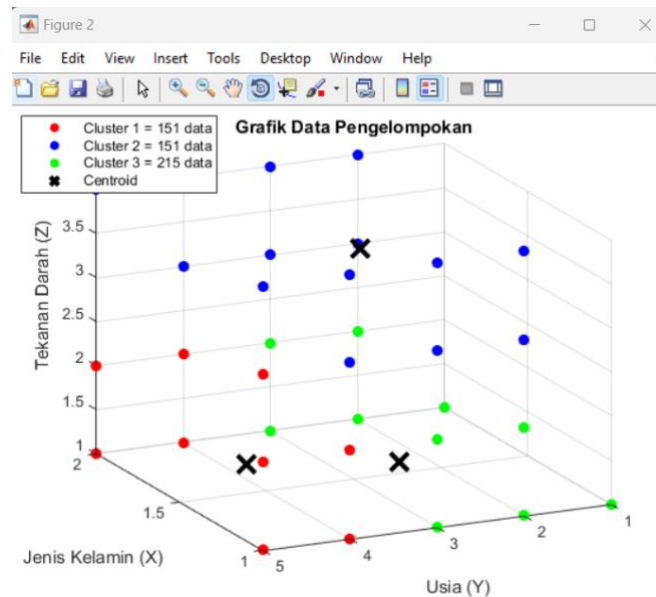


Fig. 1: Results of the three-cluster configuration

The three-cluster test divided the 517 patient records into Cluster 1 with 151 records, Cluster 2 with 151 records, and Cluster 3 with 215 records. Each point in the three-dimensional graph represented a patient record, while the centroid markers represented the average characteristics of each group.

The centroids and number of members generated from the three-cluster configuration are presented in Table 2.

Table 2: Centroids of the three-cluster configuration

Cluster	Gender centroid (X)	Age centroid (Y)	Systolic pressure centroid (Z)	Number of records
1	1.4768	4.2715	1.3642	151
2	1.4172	3.0861	3.7219	151
3	1.4326	2.6000	1.2140	215

Based on the centroid characteristics, the three clusters were interpreted as follows:

a) Cluster 1: Moderate hypertension severity.

Cluster 1 contained 151 patient records. Its age centroid of 4.2715 indicated that the group was dominated by older patients, particularly those approximately 61 years of age or older. However, its systolic blood pressure centroid of 1.3642 remained between the 140–150 mmHg and 151–160 mmHg categories. Accordingly, this cluster was interpreted as the moderate-severity group.

b) Cluster 2: High hypertension severity.

Cluster 2 contained 151 patient records. Its systolic blood pressure centroid of 3.7219 was the highest among the three groups and approached the transformation category of ≥ 171 mmHg. This characteristic indicated that Cluster 2 consisted of patients with relatively high systolic blood pressure and was therefore interpreted as the high-severity group.

c) Cluster 3: Low hypertension severity.

Cluster 3 contained 215 patient records and represented the largest group. Its age centroid of 2.6000 and systolic blood pressure centroid of 1.2140 were lower than those of the other clusters. The group was therefore interpreted as consisting of relatively younger patients with systolic blood pressure predominantly close to the 140–150 mmHg category.

Cluster variance analysis produced values of 0.7186 for Cluster 1, 1.3703 for Cluster 2, and 0.7187 for Cluster 3. Cluster 1 and Cluster 3 had relatively compact distributions, whereas Cluster 2 had the highest variance and the most heterogeneous distribution. The average cluster variance for the three-cluster configuration was 0.9359.

Table 3: Variance evaluation of the three-cluster configuration

Cluster	Number of records	Variance	Minimum distance	Maximum distance
1	151	0.7186	0.4338	3.4338
2	151	1.3703	0.2588	4.0800
3	215	0.7187	0.3929	2.9278

2. Results of four clusters

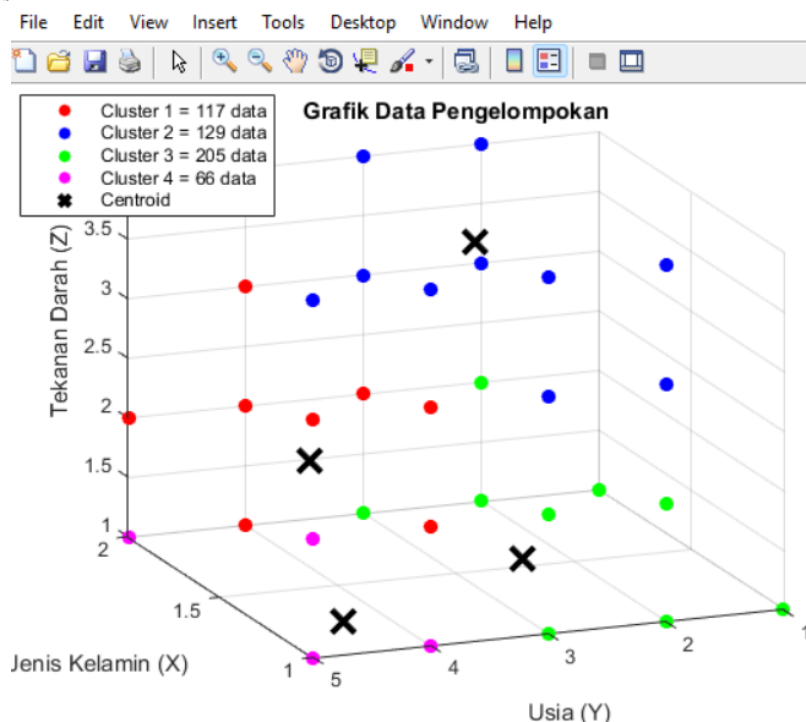


Fig. 2: Results of the four-cluster configuration

The four-cluster test was performed to determine whether the patient records could be separated into more specific groups. The results divided the dataset into Cluster 1 with 117 records, Cluster 2 with 129 records, Cluster 3 with 205 records, and Cluster 4 with 66 records. The centroid values produced by the four-cluster configuration are presented in Table 4.

Table 4: Centroids of the four-cluster configuration

Cluster	Gender centroid (X)	Age centroid (Y)	Systolic pressure centroid (Z)	Number of records
1	1.6581	3.9915	1.8889	117
2	1.4419	2.9302	3.8450	129
3	1.4049	2.5805	1.1756	205
4	1.1667	4.4848	1.0758	66

The characteristics of the four clusters were interpreted as follows:

- Cluster 1** represented patients predominantly aged approximately 61–70 years, with systolic blood pressure approaching the 151–160 mmHg category. This cluster was interpreted as a moderate-hypertension group.
- Cluster 2** had the highest systolic blood pressure centroid, at 3.8450, which approached the ≥ 171 mmHg category. This cluster was interpreted as the high-hypertension group.
- Cluster 3** contained 205 records and had relatively low age and systolic blood pressure centroids. Its systolic blood pressure centroid approached the 140–150 mmHg category, indicating a relatively low-severity group.
- Cluster 4** had the highest age centroid but the lowest systolic blood pressure centroid. It therefore represented older patients whose systolic blood pressure remained predominantly within the lower category.

The variance values were 0.9134 for Cluster 1, 1.2952 for Cluster 2, 0.6975 for Cluster 3, and 0.4587 for Cluster 4. Cluster 4 was the most homogeneous group, while Cluster 2 had the broadest distribution.

Table 5: Variance evaluation of the four-cluster configuration

Cluster	Number of records	Variance	Minimum distance	Maximum distance
1	117	0.9134	0.1293	2.6849
2	129	1.2952	0.2241	4.6195
3	205	0.6975	0.3708	2.8830
4	66	0.4587	0.2686	1.1474

The average cluster variance was 0.8412, which was lower than the value of 0.9359 obtained from the three-cluster configuration. This result indicated that four clusters produced a more compact overall grouping. However, the additional cluster separated older patients with relatively low systolic blood pressure rather than forming a new severity category.

3. Results of five clusters

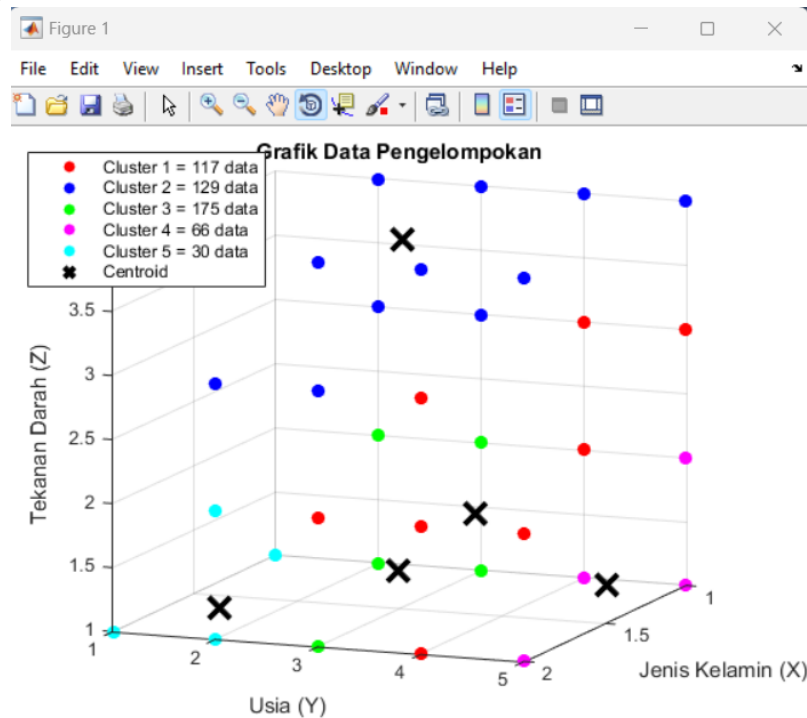


Fig. 3: Results of the five-cluster configuration

The five-cluster test produced a more detailed division of patient characteristics. Cluster 1 contained 117 records, Cluster 2 contained 129 records, Cluster 3 contained 175 records, Cluster 4 contained 66 records, and Cluster 5 contained 30 records.

The centroid values produced by this configuration are presented in Table 6.

Table 6: Centroids of the five-cluster configuration

Cluster	Gender centroid (X)	Age centroid (Y)	Systolic pressure centroid (Z)	Number of records
1	1.6581	3.9915	1.8889	117
2	1.4419	2.9302	3.8450	129
3	1.3314	2.7200	1.1829	175
4	1.1667	4.4848	1.0758	66
5	1.8333	1.7667	1.1333	30

The five clusters were interpreted as follows:

- Cluster 1 represented patients predominantly aged 61–70 years, with systolic blood pressure approaching the 151–160 mmHg category. This group was interpreted as having moderate hypertension severity.
- Cluster 2 had the highest systolic blood pressure centroid, approaching the ≥ 171 mmHg category. This group was interpreted as having high hypertension severity.
- Cluster 3 primarily represented patients approximately 51–60 years of age whose systolic blood pressure remained close to the 140–150 mmHg category. This cluster represented a lower-severity group.
- Cluster 4 represented older patients, as indicated by the highest age centroid, but its systolic blood pressure centroid was the lowest among the five groups. It therefore represented older patients with relatively low systolic blood pressure.
- Cluster 5 had the lowest age centroid and a low systolic blood pressure centroid. This group represented relatively younger patients with systolic blood pressure predominantly close to 140–150 mmHg.

The variance values were 0.9134 for Cluster 1, 1.2952 for Cluster 2, 0.5726 for Cluster 3, 0.4587 for Cluster 4, and 0.4333 for Cluster 5. Cluster 5 was the most homogeneous, whereas Cluster 2 remained the most heterogeneous.

Table 7: Variance evaluation of the five-cluster configuration

Cluster	Number of records	Variance	Minimum distance	Maximum distance
1	117	0.9134	0.1293	2.6849
2	129	1.2952	0.2241	4.6195
3	175	0.5726	0.2217	1.2960
4	66	0.4587	0.2686	1.1474
5	30	0.4333	0.1000	1.3000

The five-cluster configuration produced an average cluster variance of 0.7347, which was lower than the values produced by the three- and four-cluster configurations. This result indicated that increasing the cluster number generated more compact groups, particularly by separating low-systolic-pressure patients according to age characteristics.

2.3. Comparative evaluation of cluster configurations

The comparison of the three configurations is presented in Table 8.

Table 8: Comparison of clustering configurations

Configuration	Number of clusters	Overall cluster variance	Primary interpretation
Three clusters	3	0.9359	Low, moderate, and high severity
Four clusters	4	0.8412	Three severity patterns and a separate older-patient group
Five clusters	5	0.7347	More detailed grouping based on age and systolic blood pressure

The decreasing cluster variance indicates that increasing the number of clusters generated groups whose members were more compact around their respective centroids. Based solely on cluster compactness, the five-cluster configuration produced the best result, with the lowest overall cluster variance of 0.7347. Nevertheless, the three-cluster configuration was retained as the principal model for interpreting hypertension severity because its groups directly corresponded to the low-, moderate-, and high-severity categories established in the research objective.

The five-cluster result provides greater statistical compactness but introduces subdivisions primarily based on age differences among patients with relatively low systolic blood pressure. Consequently, it is more suitable for detailed patient segmentation, whereas the three-cluster configuration offers clearer practical interpretation for severity mapping. In all configurations, the clustering output should be understood as analytical decision-support information and not as a direct clinical diagnosis. This conclusion is consistent with the final interpretation of the thesis, which distinguishes statistical compactness from practical severity interpretation.

3. Conclusion

Based on the implementation of the K-Means algorithm using MATLAB R2014b on 517 medical records of hypertensive patients at the Anugrah Binjai Primary Care Clinic, the following conclusions were obtained:

- The K-Means algorithm successfully grouped patient data based on gender, age, and systolic blood pressure. The clustering process involved data selection, numerical transformation, initial centroid determination, Euclidean Distance calculation, data allocation to the nearest centroid, and iterative centroid updating until the cluster membership reached convergence. The results indicate that K-Means can systematically identify similarities among hypertensive patient characteristics.
- In the three-cluster configuration, Cluster 1 consisted of 151 patients and was interpreted as the moderate-severity group, Cluster 2 consisted of 215 patients and was interpreted as the low-severity group. The overall cluster variance was 0.9359. Although Cluster 2 showed the largest internal variance, this configuration provided the clearest interpretation according to the low, moderate, and high hypertension severity categories.
- In the four-cluster configuration, the patients were divided into 117 records in Cluster 1, 129 records in Cluster 2, 205 records in Cluster 3, and 66 records in Cluster 4. The overall cluster variance decreased to 0.8412, indicating a more compact grouping than the three-cluster configuration. Cluster 4 had the smallest variance and represented older patients with relatively low systolic blood pressure, while Cluster 2 remained the group with the highest systolic blood pressure and the broadest data distribution.
- In the five-cluster configuration, Cluster 1 consisted of 117 records, Cluster 2 of 129 records, Cluster 3 of 175 records, Cluster 4 of 66 records, and Cluster 5 of 30 records. The overall cluster variance was 0.7347, which was the lowest among the three configurations. Cluster 5 had the smallest individual variance of 0.4333, whereas Cluster 2 had the largest variance of 1.2952. These findings indicate that the five-cluster configuration produced the most compact and homogeneous grouping.

Overall, the five-cluster configuration provided the best statistical clustering quality because it produced the lowest overall cluster variance. However, the three-cluster configuration was selected as the principal model for interpreting hypertension severity because its clusters directly corresponded to the low, moderate, and high categories established in the research objective. The clustering results can support the clinic in mapping patient characteristics, planning targeted health education, and prioritizing patient monitoring. Nevertheless, the resulting clusters should be used as data-driven decision-support information and not as a substitute for clinical diagnosis by healthcare professionals.

References

- [1] B. Laksono, Y. Syahidin, and Y. Yunengsih, "Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering," *Jurnal Teknologi Sistem Informasi dan Aplikasi*, vol. 7, no. 2, pp. 621–627, Apr. 2024, doi: 10.32493/jtsi.v7i2.39354.
- [2] S. Audia, R. Buaton, and Z. Fatmaira, "Journal of Artificial Intelligence and Engineering Applications Clustering of High-Achieving Students Based on Scores at Junior High School Level Using K-Means Algorithm," 2025. [Online]. Available: <https://ioinformatic.org/>
- [3] U. Hayati, "CLUSTERING PRODUK EKSPOR INDONESIA BERDASARKAN TINGKAT PERMINTAAN MENGGUNAKAN METODE K-MEANS TAHUN 2020-2022," 2023.
- [4] I. A. Sifai and R. Wulandari, "Prevalensi dan Faktor Risiko Hipertensi pada Orang Dewasa di Semarang Barat," *MEDIA KESEHATAN MASYARAKAT INDONESIA*, vol. 23, no. 4, pp. 344–350, Dec. 2024, doi: 10.14710/mkmi.23.4.344-350.
- [5] M. G. Rohman, K. Yahya, and P. H. Susilo, "Implementasi Algoritma K-means Clustering pada Pengelompokan Data Kepuasan Penggunaan E-Learning," *Generation Journal*, vol. 8, no. 2, pp. 2580–4952, Aug. 2024.

- [6] Y. Tedju, Y. R. Kaesmetan, and M. I. Lamabelawa, "PENGELOMPOKAN PENYAKIT HIPERTENSI DI KOTA KUPANG MENGGUNAKAN METODE K-MEANS," *JURNAL TEKNOLOGI INFORMASI*, 2021.
- [7] U. Qalsum and W. Abidin, "Klasifikasi Penyakit Hipertensi Menggunakan Metode K-Means Clustering," *Jurnal Matematika dan Statistika serta Aplikasinya*, vol. 11, no. 2, p. 2023, 2023.
- [8] J. Wala and R. Umar, "Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung," 2024.
- [9] R. Anggita, R. Buaton, K. Annatasia Br Sitepu, S. Informasi, and S. Kaputama, "Pengelompokan Data Keluhan Pasien Pada Layanan Rumah Sakit Berdasarkan Kategori Masalah Menggunakan Metode Clustering," *Desember*, vol. 6, no. 2, 2025.
- [10] W. Aulia, A. Putera Utama Siahaan, L. Marlina, and M. Iqbal, "ANALISIS ALGORITMA K-MEANS CLUSTERING DALAM IDENTIFIKASI TINGKAT RISIKO PENYAKIT BERDASARKAN DATA REKAM MEDIS PASIEN," 2025. [Online]. Available: <http://jurnal.goretanpena.com/index.php/JSSR>
- [11] S. Kosamah, F. Aulia Lubis, M. Faris Al Rafiq, and Z. Putri Julia Daulay, "Clustering Pola Intensitas Curah Hujan Berbasis Data Radar NOAA NEXRAD Level II Menggunakan Algoritma K-Means," *Jurnal Ilmu Teknik dan Informatika*, vol. 6, pp. 45–59, May 2026, doi: 10.51903/teknik.v6i1.1135.
- [12] Dea Syafitri, Yani Maulita, and Lina Arlianan Nur Kadim, "Pengelompokan Data Keluhan Masyarakat Terhadap Fasilitas Umum di Kota Binjai Menggunakan Metode Clustering," *Jurnal Penelitian Teknologi Informasi dan Sains*, vol. 2, no. 3, pp. 158–170, Aug. 2024, doi: 10.54066/jptis.v2i3.2381.