INCODING: Journal of Informatic and Computer Science Engineering
https://journal.mahesacenter.org/index.php/incoding/index || ISSN: 2776-432X (Online)
5 (2) 2025: 250-268

DOI: 10.34007/incoding.v5i2.973

Penerapan Algoritma K-Means Untuk Klasterisasi Pasien Berdasarkan Riwayat Kesehatan dan Jenis Layanan Kesehatan
Application of K-Means Algorithm for Clustering of Patients Based on Health History and Type of Health Service

Riza Dwi Putri & Rizki Muliono*
Program Studi Teknik Informatika, Universitas Medan Area, Indonesia
Diterima: 16 April 2025; Direview: 20 April 2025; Disetujui: 24 April 2025
Email: rizimuliono@staff.uma.ac.id

Abstrak

Transformasi digital dalam dunia kesehatan telah menghasilkan volume data yang besar dan kompleks, sehingga memerlukan teknik analisis yang tepat untuk menghasilkan informasi bermakna. Penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam mengelompokkan pasien berdasarkan riwayat kesehatan dan jenis layanan kesehatan yang digunakan, guna mendukung pengambilan keputusan manajemen rumah sakit secara data-driven. Data yang digunakan berasal dari 1.459 pasien Rumah Sakit Sapta Medika yang mencakup atribut usia, jenis kelamin, riwayat penyakit kronis (diabetes, hipertensi, jantung), jumlah kunjungan, biaya pengobatan, serta jenis layanan seperti rawat jalan, rawat inap, IGD, dan telemedis. Tahapan penelitian dimulai dari pra-pemrosesan data, transformasi, encoding data kategorikal, normalisasi data numerik, hingga proses klasterisasi dengan K-Means. Metode Elbow digunakan untuk menentukan jumlah klaster optimal, yang diperoleh pada K = 3. Hasil klasterisasi menunjukkan terbentuknya tiga klaster dengan karakteristik berbeda, yaitu: klaster pasien kronis dengan biaya tinggi dan frekuensi rawat inap yang sering, klaster pasien rutin yang stabil menggunakan rawat jalan, serta klaster pasien umum yang cenderung muda dengan kasus ringan. Visualisasi menggunakan PCA memperlihatkan pemisahan klaster yang baik, sedangkan evaluasi menggunakan Silhouette Score mencapai nilai 0,47. Hasil ini menyimpulkan bahwa algoritma K-Means mampu menghasilkan segmentasi pasien yang representatif dan aplikatif, serta dapat menjadi dasar dalam perancangan layanan kesehatan yang adaptif, efisien, dan berbasis kebutuhan pasien.
Kata Kunci: Klasterisasi Pasien; Data Kesehatan; K-Means; Layanan Kesehatan; Segmentasi Medis

Abstract

The digital transformation in the healthcare sector has led to the generation of large and complex datasets, requiring appropriate analytical techniques to extract meaningful information. This study aims to implement the K-Means algorithm to cluster patients based on their health history and the types of healthcare services they use, in order to support data-driven decision-making in hospital management. The dataset consists of 1,459 patient records from Sapta Medika Hospital, covering attributes such as age, gender, chronic disease history (diabetes, hypertension, heart disease), visit frequency, medical costs, and healthcare service types including outpatient, inpatient, emergency (ER), and telemedicine. The research stages involved data preprocessing, transformation, categorical data encoding, numerical data normalization, and clustering using the K-Means algorithm. The optimal number of clusters was determined using the Elbow Method, which identified K = 3. The clustering results revealed three distinct patient groups: chronic patients with high treatment costs and frequent inpatient services, routine patients with stable conditions mostly using outpatient services, and general patients, usually younger with mild conditions. Principal Component Analysis (PCA) was used to visualize the cluster separation, while the clustering quality was evaluated with a Silhouette Score of 0.47. These results conclude that the K-Means algorithm is effective in producing meaningful and practical patient segmentation, which can be used to design more adaptive, efficient, and patient-centered healthcare service strategies.
***Keywords**: Patient Clustering; Health Data; K-Means; Healthcare Services; Medical Segmentation*

PENDAHULUAN

Dalam beberapa tahun terakhir, transformasi digital di sektor kesehatan telah mengalami percepatan yang signifikan, baik dari sisi pencatatan data pasien secara elektronik, pengembangan layanan berbasis telemedis, hingga penggunaan kecerdasan buatan dalam mendukung pengambilan keputusan klinis [1]. Salah satu dampak dari transformasi ini adalah meningkatnya volume data kesehatan yang tersedia dan perlu dianalisis secara efektif [2]. Namun, data yang besar dan kompleks tidak akan memberi manfaat maksimal tanpa adanya teknik analisis yang tepat untuk menggali informasi bermakna di dalamnya. Oleh karena itu, pendekatan data mining, khususnya teknik klasterisasi, menjadi semakin relevan dalam konteks ini, terutama untuk mengidentifikasi pola-pola tersembunyi di antara pasien berdasarkan karakteristik medis dan layanan kesehatan yang mereka gunakan [3][4]. Salah satu algoritma klasterisasi yang paling banyak digunakan dalam literatur adalah algoritma K-Means, yang memiliki keunggulan dalam hal kecepatan, kemudahan implementasi, dan interpretasi hasil yang intuitif [5][6][7]. Dalam dunia medis, algoritma ini telah digunakan untuk berbagai keperluan seperti segmentasi pasien, pengelompokan gejala penyakit, dan analisis perilaku Kesehatan [8][9][10]. Misalnya, Zhang et al. (2020) berhasil menerapkan K-Means untuk mengelompokkan pasien rumah sakit berdasarkan data rekam medis elektronik (electronic health records/EHR), dan menunjukkan bahwa metode ini mampu mendeteksi pola risiko penyakit kronis dengan akurasi yang baik. Segmentasi pasien berdasarkan fitur klinis dan gaya hidup dapat membantu meningkatkan efektivitas strategi intervensi medis [11][8]. Bahkan dalam konteks layanan darurat menunjukkan bahwa klasterisasi pasien berdasarkan perilaku kunjungan ke IGD (Instalasi Gawat Darurat) dapat mengoptimalkan alokasi sumber daya medis dan mempercepat proses triase [12][8].

Namun demikian, hasil penelusuran terhadap berbagai penelitian mutakhir menunjukkan adanya beberapa keterbatasan atau kesenjangan (research gap) yang belum banyak disoroti secara mendalam. Pertama, sebagian besar studi sebelumnya berfokus pada parameter klinis atau data biologis pasien, seperti hasil laboratorium, tekanan darah, atau kadar gula darah [13][10][14], dan belum banyak yang menggabungkan antara data riwayat penyakit kronis dengan jenis layanan kesehatan yang digunakan pasien, seperti rawat jalan, rawat inap, telemedis, atau IGD. Padahal, jenis layanan yang dipilih pasien dapat mencerminkan perilaku pencarian layanan kesehatan (health-seeking behavior)

serta kondisi kesehatannya secara umum. Misalnya, pasien dengan penyakit kronis yang rutin menggunakan layanan rawat jalan akan memiliki profil yang berbeda dibandingkan pasien yang jarang berobat namun sering ke IGD [12][15][3]. Kedua, sebagian besar penelitian yang ada hanya menggunakan data dari satu institusi kesehatan. Hal ini berpotensi menghasilkan bias institusional karena karakteristik pasien di satu rumah sakit bisa sangat berbeda dengan rumah sakit lainnya, baik dari segi demografi, jenis penyakit dominan, maupun akses terhadap layanan Kesehatan [11][3]. Studi yang tidak mempertimbangkan variasi antar rumah sakit akan sulit digeneralisasikan untuk konteks yang lebih luas. Oleh karena itu, penting untuk mempertimbangkan data dari berbagai rumah sakit untuk melakukan simulasi, agar hasil segmentasi pasien dapat mencerminkan realitas sistem kesehatan yang lebih kompleks [4][16]. Ketiga, dalam konteks sistem informasi manajemen rumah sakit, sangat sedikit penelitian yang mencoba menyatukan atribut demografi, status penyakit, frekuensi kunjungan, dan jenis layanan ke dalam satu model segmentasi pasien secara utuh [17][18]. Padahal pendekatan seperti ini sangat potensial untuk dikembangkan sebagai decision support system yang mampu memberi rekomendasi layanan atau intervensi berdasarkan klaster pasien tertentu [8][3]. Misalnya, pasien dari klaster tertentu yang berisiko tinggi dapat diarahkan untuk mengikuti program pemantauan kesehatan rutin, sementara klaster lain mungkin hanya memerlukan edukasi preventif [10].

Berdasarkan paparan tersebut, penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam mengelompokkan pasien berdasarkan kombinasi atribut demografi (umur, jenis kelamin), riwayat penyakit kronis (diabetes, hipertensi, jantung), intensitas layanan (frekuensi kunjungan, jumlah diagnosa, riwayat rawat inap), jenis layanan yang digunakan (rawat jalan, rawat inap, IGD, telemedis), serta sumber rumah sakit [1]. Penggunaan atribut gabungan ini diharapkan dapat menghasilkan segmentasi pasien yang lebih komprehensif dan aplikatif dalam konteks sistem layanan kesehatan modern.

Metode Elbow digunakan dalam penelitian ini untuk menentukan jumlah klaster optimal dalam algoritma K-Means. Metode ini memanfaatkan Inertia atau jumlah kuadrat jarak antar titik data dengan pusat klaster untuk mengevaluasi kualitas klaster yang terbentuk [21]. Dengan menganalisis grafik Inertia terhadap jumlah klaster, titik elbow akan menunjukkan jumlah klaster yang optimal, yaitu titik dimana penurunan Inertia

mulai melambat secara signifikan. Hal ini memberikan gambaran yang jelas tentang batas yang tepat antara peningkatan klaster yang memberikan hasil signifikan dan yang hanya menambah kompleksitas tanpa memperbaiki hasil klasterisasi. (Kaufman & Rousseeuw, 1990) mengemukakan dalam artikel mereka bahwa metode Elbow merupakan salah satu cara yang sederhana namun efektif dalam memilih jumlah klaster yang tepat. Selain itu, untuk evaluasi hasil klasterisasi, Siluet Score juga digunakan untuk mengukur kualitas klaster. Siluet Score menggabungkan dua aspek utama dalam klasterisasi: kedekatan antar titik dalam klaster yang sama dan jarak antar klaster yang berbeda. Nilai Siluet berkisar antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan klaster yang lebih baik, karena data yang lebih terpisah dengan jelas [22]. Dalam penelitian ini, Siluet Score digunakan untuk memastikan bahwa klaster yang terbentuk adalah homogen dan terpisah dengan jelas satu sama lain, memberikan indikasi bahwa hasil klasterisasi dapat diandalkan.

Tujuan spesifik dari penelitian ini adalah untuk mengembangkan model klasterisasi pasien berbasis algoritma K-Means dengan mempertimbangkan berbagai fitur multivariabel, seperti data rekam medis dan jenis layanan yang digunakan. Penelitian ini bertujuan untuk mengidentifikasi karakteristik masing-masing klaster yang terbentuk, sehingga dapat mendukung pengambilan keputusan berbasis data di lingkungan rumah sakit. Selanjutnya, penelitian ini akan menilai perbedaan karakteristik pasien antar rumah sakit berdasarkan hasil klasterisasi, yang nantinya dapat dijadikan dasar untuk perencanaan layanan dan kebijakan berbasis wilayah. Dengan pendekatan ini, penelitian diharapkan memberikan kontribusi signifikan pada pengembangan sistem pendukung keputusan dalam manajemen layanan kesehatan, melalui

segmentasi pasien yang lebih adaptif dan berbasis data multisumber. Di akhir, penelitian ini diharapkan tidak hanya memberikan kontribusi ilmiah dalam hal metodologi dan penerapannya, tetapi juga memberikan kontribusi praktis dalam pengembangan sistem informasi kesehatan, mendukung personalisasi layanan medis, serta mendorong transformasi manajemen rumah sakit menuju arah yang lebih berbasis data.

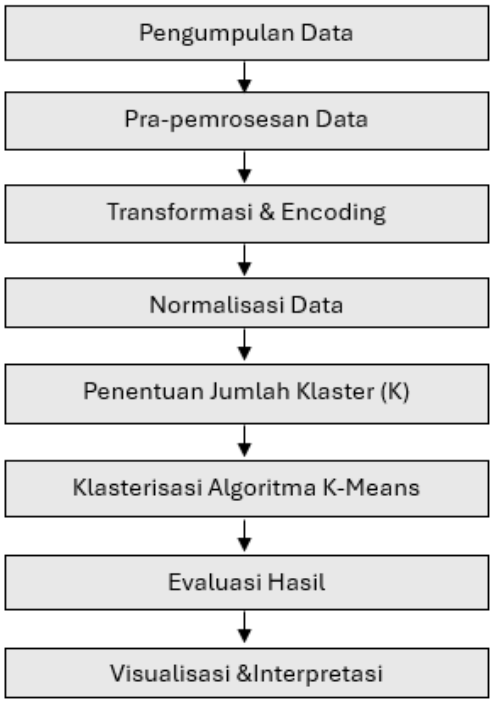
METODE PENELITIAN

Jenis Penelitian.

Penelitian ini menggunakan pendekatan kuantitatif deskriptif dengan metode eksperimen berbasis algoritma K-Means untuk melakukan klasterisasi pasien berdasarkan atribut riwayat kesehatan dan jenis layanan kesehatan. Tujuan utama dari penelitian ini adalah untuk mengelompokkan pasien ke dalam beberapa klaster homogen agar diperoleh pemahaman yang lebih baik terkait profil masing-masing segmen pasien. Penelitian dilakukan dengan menggunakan dataset sebanyak 1459 data pasien dari rumah sakit Sapta Medika yang disimulasikan.

Tahapan Penelitian.

Tahapan pelaksanaan penelitian secara umum digambarkan pada Gambar 1 berikut ini:



Gambar 1. Tahapan penelitian

Gambar 1 menggambarkan tahapan sistematis dalam pelaksanaan penelitian klasterisasi data pasien menggunakan algoritma K-Means. Tahapan pertama dimulai dengan pengumpulan data, yang berasal dari rekam medis pasien dan mencakup atribut-atribut seperti umur, jenis kelamin, riwayat penyakit kronis, frekuensi kunjungan, biaya pengobatan, jenis layanan yang digunakan (rawat jalan, rawat inap, IGD, atau telemedis),

serta tipe pembiayaan (BPJS, mandiri, atau asuransi swasta). Setelah data terkumpul, langkah selanjutnya adalah pra-pemrosesan data, yaitu proses pembersihan data dari duplikasi, penanganan nilai kosong, serta standarisasi format data agar siap digunakan untuk analisis lebih lanjut. Setelah pra-pemrosesan, dilakukan transformasi dan encoding terhadap atribut kategorikal, seperti Jenis_Kelamin dan Jenis_Layanan, agar dapat diubah menjadi format numerik menggunakan teknik One-Hot Encoding. Data yang telah dikonversi kemudian dinormalisasi menggunakan metode Min-Max Scaling untuk memastikan bahwa semua fitur memiliki rentang nilai yang seragam dan tidak mendominasi perhitungan jarak antar data. Tahap berikutnya adalah penentuan jumlah klaster optimal (K) dengan menggunakan metode Elbow, yang mengamati titik tekuk dari grafik WCSS sebagai dasar pemilihan jumlah klaster. Proses klasterisasi kemudian dilakukan menggunakan algoritma K-Means dengan parameter K yang telah ditentukan. Data dikelompokkan ke dalam beberapa klaster berdasarkan kemiripan karakteristik antar pasien. Setelah proses klasterisasi selesai, hasilnya dievaluasi menggunakan metrik Silhouette Score, yang mengukur sejauh mana anggota dalam suatu klaster memiliki kesamaan dan berbeda secara signifikan dengan klaster lain. Tahap terakhir adalah visualisasi dan interpretasi hasil, di mana hasil klasterisasi divisualisasikan menggunakan teknik reduksi dimensi PCA (Principal Component Analysis), dan masing-masing klaster dianalisis untuk memahami profil serta kebutuhan layanan kesehatan dari setiap kelompok pasien.

Pengumpulan Dan Persiapan Data.

Data yang digunakan dalam penelitian ini diperoleh dari rekam medis pasien di Rumah Sakit Sapta Medika selama periode tertentu. Data diseleksi dan dianonimkan terlebih dahulu untuk menjaga privasi pasien sesuai dengan prinsip etika penelitian. Jumlah data yang dianalisis adalah sebanyak 1459 pasien yang terdiri dari berbagai jenis layanan dan kategori penyakit. Atribut data pasien yang digunakan disajikan dalam **Tabel 1**

Tabel 1 Atribut Data Pasien RS Sapta Medika

No	Nama Atribut	Jenis Data	Keterangan
1	ID_Pasien	String	Identifikasi unik pasien (disamarkan)
2	Umur	Numerik	Usia pasien
3	Jenis_Kelamin	Kategorikal	Pria atau Wanita
4	Diabetes	Biner (0/1)	Riwayat penyakit diabetes
5	Hipertensi	Biner (0/1)	Riwayat hipertensi

No	Nama Atribut	Jenis Data	Keterangan
6	Jantung	Biner (0/1)	Riwayat penyakit jantung
7	Kunjungan_Tahunan	Numerik	Frekuensi kunjungan ke rumah sakit dalam satu tahun
8	Riwayat_Rawat_Inap	Biner (0/1)	Indikasi pernah dirawat inap
9	Jumlah_Diagnosa	Numerik	Banyaknya penyakit yang didiagnosis
10	Biaya_Total_Pengobatan	Numerik	Estimasi total biaya pengobatan (Rp)
11	Jenis_Layanan	Kategorikal	Rawat Jalan, Rawat Inap, IGD, atau Telemedis
12	Tipe_Pembiayaan	Kategorikal	BPJS, Mandiri, atau Asuransi Swasta

Prapemrosesan Dan Transformasi Data.

Sebelum diterapkannya algoritma K-Means, data pasien dari Rumah Sakit Sapta Medika terlebih dahulu melalui proses pra-pemrosesan dan transformasi untuk memastikan bahwa data dalam kondisi bersih, konsisten, dan sesuai untuk dianalisis secara numerik. Tahapan ini sangat penting karena kualitas data sangat memengaruhi hasil akhir dari proses klasterisasi.

1. Pembersihan data: menghapus duplikat dan memperbaiki data kosong atau tidak valid.
2. Encoding data kategorikal menggunakan One-Hot Encoding agar dapat digunakan dalam model K-Means.
3. Normalisasi fitur numerik dengan teknik Min-Max Scaling, sehingga semua atribut berada dalam rentang [0, 1] dan tidak mendominasi perhitungan jarak dalam algoritma.

Penentuan Jumlah Klaster (K).

Penentuan jumlah klaster optimal dilakukan menggunakan metode Elbow dan Silhouette Score. Metode Elbow dihitung melalui *Within-Cluster Sum of Squares* (WCSS) untuk berbagai nilai KKK dengan rumus:

$$\mathrm{WCSS}(K) = \sum_{c=1}^{K} \sum_{x_i \in C_c} \|x_i - \mu_c\|^2$$

di mana X_1 adalah titik data, μ_c adalah centroid klaster c, dan C_c adalah himpunan data dalam klaster tersebut. Nilai WCSS akan menurun seiring bertambahnya ᴋ, dan titik "siku" pada grafik menandakan jumlah klaster yang efisien.

Metode Silhouette Score digunakan untuk mengukur kualitas pemisahan klaster. Untuk setiap data i, dihitung:

A(i): rata-rata jarak ke semua titik dalam klaster yang sama.

B(i): rata-rata jarak ke semua titik di klaster terdekat lainnya.

Rumus Silhouette:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Nilai s(i) berada antara −1 hingga 1, dengan nilai mendekati 1 menunjukkan pemisahan yang baik antar klaster. Nilai rata-rata Silhouette digunakan sebagai indikator kualitas klasterisasi.

Berdasarkan kedua metode tersebut, jumlah klaster optimal adalah K = 3, karena grafik Elbow menunjukkan titik siku yang jelas pada K=3 dan nilai Silhouette rata-rata sebesar ≈ 0,47 menunjukkan kualitas pemisahan yang memadai.

Penerapan Algoritma K-Means.

Algoritma K-Means digunakan untuk membagi data ke dalam jumlah klaster KKK yang telah ditentukan. Prosesnya meliputi:

1. Menentukan jumlah klaster K (hasil dari tahap sebelumnya).
2. Menginisialisasi centroid secara acak.
3. Menghitung jarak setiap data X_i ke setiap centroid menggunakan jarak Euclidean:

$$d(x_i, \mu_c) = \sqrt{\sum_{j=1}^{m} (x_{ij} - \mu_{cj})^2}$$

4. Menetapkan data ke klaster dengan jarak terkecil.
5. Memperbarui centroid dengan menghitung rata-rata semua data dalam klaster:

$$\mu_c = \frac{1}{|C_c|} \sum_{x_i \in C_c} x_i$$

Mengulangi langkah 3–5 hingga centroid stabil atau jumlah iterasi maksimum tercapai.

Evaluasi dan Analisis Klaster.

Hasil klasterisasi pasien dievaluasi menggunakan:

1. Silhouette Score, untuk mengukur koherensi dalam klaster dan tingkat pemisahan antar klaster. Nilai rata-rata ≈ 0,47 menunjukkan bahwa klaster terbentuk dengan pemisahan yang memadai.
2. Profil statistik tiap klaster, meliputi rata-rata umur pasien, proporsi pasien dengan penyakit kronis, jenis layanan kesehatan yang dominan, dan rata-rata biaya layanan. Profil ini memberikan gambaran karakteristik unik setiap klaster.
3. Visualisasi klaster menggunakan metode *Principal Component Analysis* (PCA) atau *t-distributed Stochastic Neighbor Embedding* (t-SNE) untuk memproyeksikan hasil klasterisasi ke dalam ruang dua dimensi.

Visualisasi ini membantu mempermudah interpretasi hasil, memvalidasi pemisahan antar klaster, serta menjadi dasar dalam merancang strategi pelayanan yang sesuai dengan karakteristik masing-masing klaster.

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan data rekam medis 1.459 pasien yang terdiri dari informasi demografis, riwayat kunjungan, biaya pengobatan, serta jenis layanan dan penyakit yang diderita. Data diperoleh dari sistem informasi rumah sakit yang telah melalui tahap validasi dan pembersihan.

Deskripsi dataset Mentah (data awal).

Data pasien dengan 12 atribut, di antaranya: identitas pasien (disamarkan), informasi demografis, riwayat penyakit, intensitas layanan kesehatan, jenis layanan, dan tipe pembiayaan. Contoh data mentah dapat dilihat pada Tabel berikut:

Tabel 2 Data Mentah

ID_ Pasien	Umur	Jenis Kelamin	Diabetes	Hipertensi	Jantung	Kunjungan	Jenis Layanan	Riwayat rawat inap	Jumlah diagnosa	Biaya_total_pengobatan	Tipe pembiayaan
P0001	69	Wanita	0	1	0	2	Rawat Jalan	1	5	1991443	Mandiri
P0002	32	Wanita	0	0	0	1	Rawat Jalan	1	5	2383074	BPJS
P0003	78	Wanita	1	1	1	2	Igd	0	4	1505346	Mandiri
P0004	38	Wanita	1	0	0	1	Igd	0	3	2249355	BPJS
P0005	41	Pria	0	0	0	5	Rawat Inap	0	3	3003066	BPJS

Selanjutnya, data ini akan diproses agar dapat digunakan untuk algoritma klasterisasi.

Pra-pemrosesan dan Transformasi Data.

Sebelum diterapkan algoritma K-Means, dilakukan tahapan pra-pemrosesan data agar dataset siap untuk analisis numerik:

1. Pembersihan Data
   a. Data duplikat dan nilai kosong dihapus.
   b. Kolom ID_Pasien dihapus karena tidak relevan untuk proses klasterisasi.
   c. Tipe data seperti Biaya_Total_Pengobatan dan Kunjungan_Tahunan diubah menjadi numerik.
2. Transformasi Kategorikal

Kolom Jenis_Kelamin, Jenis_Layanan, dan Tipe_Pembiayaan diubah menjadi format numerik dengan metode One-Hot Encoding.

Tabel 3 Hasil encoding

Jenis_Kelamin_Wanita	Jenis_Layanan_Rawat Inap	Jenis_Layanan_Rawat Jalan	Jenis_Layanan_Telemedis	Tipe_Pembiayaan_BPJS	Tipe_Pembiayaan_Mandiri
1	0	1	0	0	1

3. Normalisasi Data

Semua fitur numerik (umur, biaya, jumlah diagnosa, dsb.) dinormalisasi menggunakan **Min-Max Scaling** ke rentang [0, 1].

Hasil Normalisasi dan Encoding:

Tabel 4 Hasil Normalisasi

Umur	Diabetes	Hipertensi	...	Jenis_Layanan_Rawat Inap	Tipe_Pembiayaan_BPJS	...
0.77	0	1	...	0	0	...
0.21	0	0	...	0	1	...
0.90	1	1	...	0	0	...

Sebelum dilakukan proses klasterisasi, data pasien yang terdiri dari atribut numerik dan kategorikal terlebih dahulu diolah agar sesuai dengan format masukan yang dibutuhkan oleh algoritma K-Means. Atribut numerik seperti Umur, Kunjungan_Tahunan, Jumlah_Diagnosa, dan Biaya_Total_Pengobatan dinormalisasi menggunakan metode **Min-Max Scaling**, yang mengubah nilai ke dalam rentang 0 hingga 1. Tujuannya adalah agar

tidak ada fitur yang mendominasi proses perhitungan jarak antar data pada algoritma K-Means, karena perbedaan skala yang terlalu jauh dapat memengaruhi hasil klasterisasi.

Sementara itu, untuk atribut kategorikal seperti Jenis_Kelamin, Jenis_Layanan, dan Tipe_Pembiayaan, dilakukan proses transformasi menggunakan teknik One-Hot Encoding. Teknik ini mengubah nilai kategorikal menjadi representasi biner (0 atau 1) untuk setiap kategori. Misalnya, atribut Jenis_Layanan yang awalnya memiliki nilai seperti

"Rawat Inap", "Rawat Jalan", "IGD", dan "Telemedis", diubah menjadi empat kolom biner yang masing-masing menunjukkan apakah pasien menggunakan layanan tersebut atau tidak.

Proses ini menjadi krusial karena kualitas hasil klasterisasi sangat bergantung pada ketepatan dan konsistensi representasi data yang digunakan. Dengan Langkah -langkah ini, data telah siap untuk memasuki tahap penentuan jumlah klaster dan proses klasterisasi menggunakan K-Means.

Tabel 5 Statistik Deskriptif Data Pasien

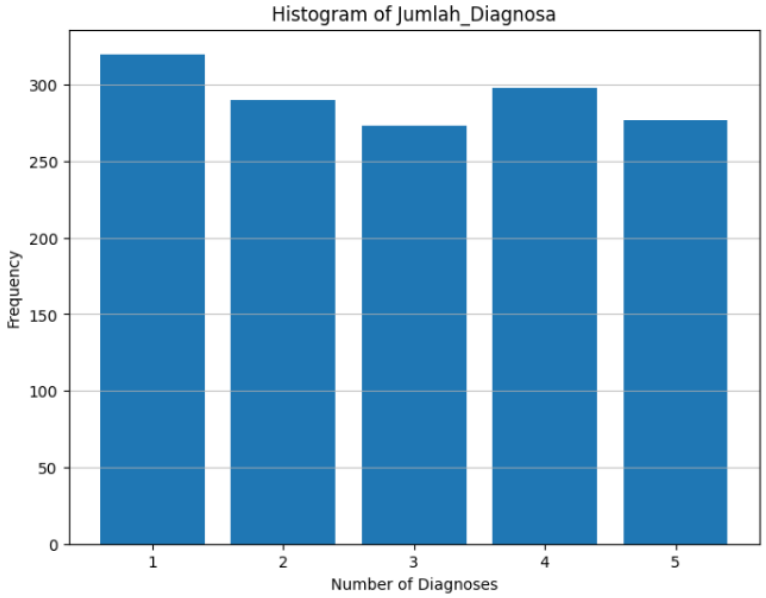
Fitur	Rata - Rata	Median	Min	Max
Umur (Tahun)	50.72	50	18	84
Kunjungan Tahunan	6.50	6	1	12
Biaya Pengobatan	Rp 4.984.333	Rp 4.865.170	Rp 104.011	Rp 9.999.007

Tabel 6 Distribusi Kategorikal

Jenis Kelamin	Jumlah
Pria	59.4%
Wanita	40.6%

Tabel 7 Jenis Layanan Kesehatan

Jenis Layanan Kesehatan	Jumlah
Rawat Inap	34.4%
Rawat Jalan	32.9%
IGD	32.7%

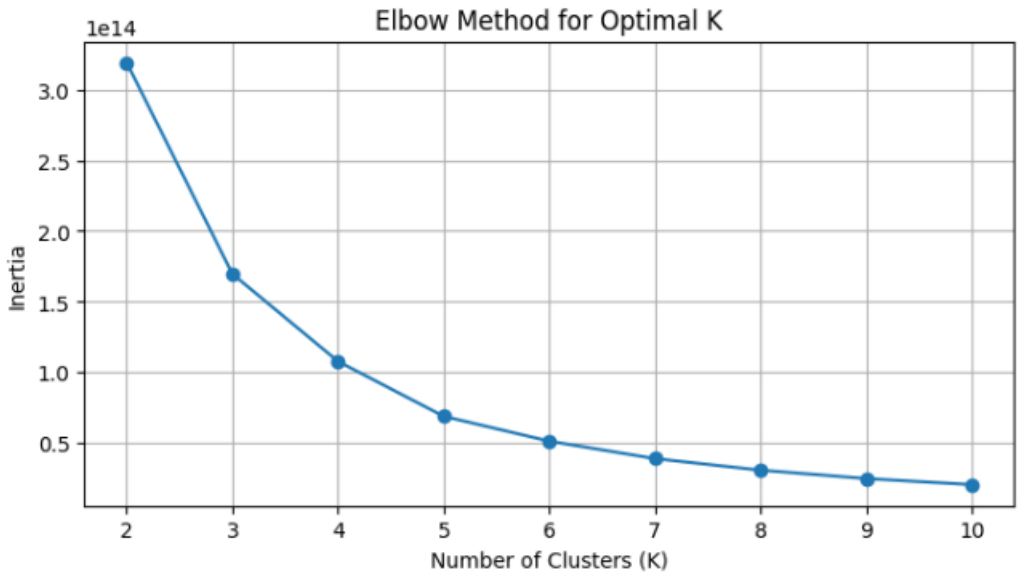


Gambar 2 Histogram Jumlah Diagnosa

Gambar 2 tersebut memperlihatkan histogram distribusi Jumlah Diagnosa yang dimiliki pasien. Sumbu horizontal menunjukkan jumlah diagnosa dari 1 hingga 5, sedangkan sumbu vertikal menunjukkan frekuensi pasien pada masing-masing kategori. Terlihat bahwa pasien dengan 1 diagnosa merupakan kelompok terbanyak, dengan frekuensi lebih dari 320 pasien. Sementara itu, pasien dengan 2 hingga 5 diagnosa memiliki jumlah yang relatif seimbang, berkisar antara 270 hingga 300 pasien. Pola distribusi ini menunjukkan bahwa mayoritas pasien hanya memiliki satu diagnosa utama, namun tidak sedikit pula pasien yang memiliki lebih dari satu diagnosa. Informasi ini penting untuk memahami kompleksitas kasus pasien, di mana kelompok dengan banyak diagnosa biasanya membutuhkan penanganan medis yang lebih intensif dan berpotensi mempengaruhi pola klasterisasi.

Hasil Klasterisasi Menggunakan K-Means.

Jumlah klaster optimal ditentukan menggunakan Elbow Method, dengan mengamati grafik *Within Cluster Sum of Squares* (WCSS) dari nilai K=2 hingga K=10. Dari grafik tersebut, titik siku (elbow point) terlihat pada K=3, sehingga dipilih 3 klaster sebagai jumlah optimal.

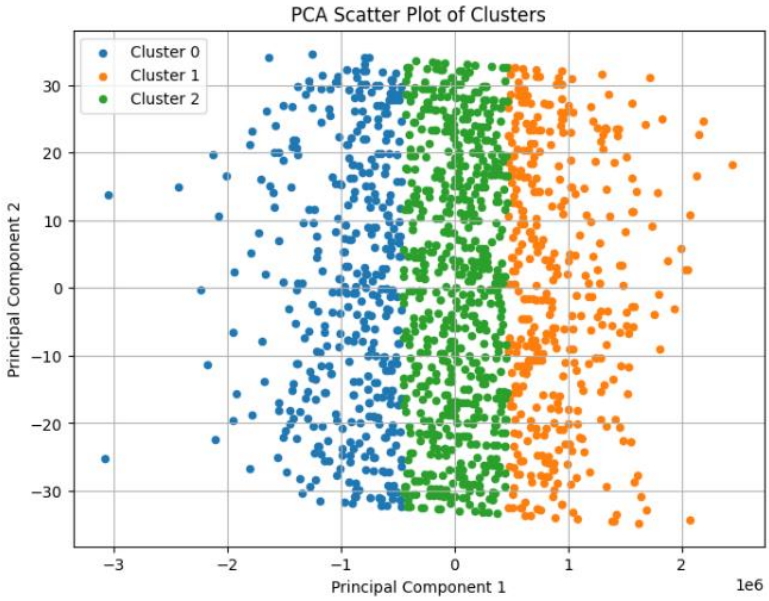


Gambar 3. metode Elbow

Gambar 3 merupakan visualisasi dari metode Elbow Gambar tersebut menunjukkan hasil penerapan Metode Elbow untuk menentukan jumlah klaster optimal pada algoritma K-Means. Sumbu horizontal menggambarkan jumlah klaster (KKK) dari 2 hingga 10, sedangkan sumbu vertikal menunjukkan nilai *inertia* atau *Within-Cluster Sum of Squares* (WCSS). Terlihat bahwa nilai inertia menurun tajam dari K=2 hingga K=3, kemudian penurunannya semakin landai pada nilai K selanjutnya. Pola ini membentuk sudut siku (elbow) yang jelas pada K=3. Titik siku tersebut mengindikasikan bahwa penggunaan lebih dari tiga klaster tidak lagi memberikan pengurangan inertia yang signifikan, sehingga jumlah klaster yang paling efisien adalah K = 3. Dengan demikian, hasil analisis Metode Elbow ini mendukung pemilihan tiga klaster sebagai jumlah klaster optimal untuk dataset yang digunakan..

Visualisasi Klaster (PCA)

Untuk memahami hasil klasterisasi secara visual, dilakukan reduksi dimensi menggunakan teknik Principal Component Analysis (PCA). PCA merupakan metode statistik yang mengubah sejumlah variabel yang saling berkorelasi menjadi sejumlah variabel baru yang tidak berkorelasi (principal components), sehingga memungkinkan visualisasi data berdimensi tinggi dalam dua dimensi.



Gambar 4 Visualisasi Klaster PCA

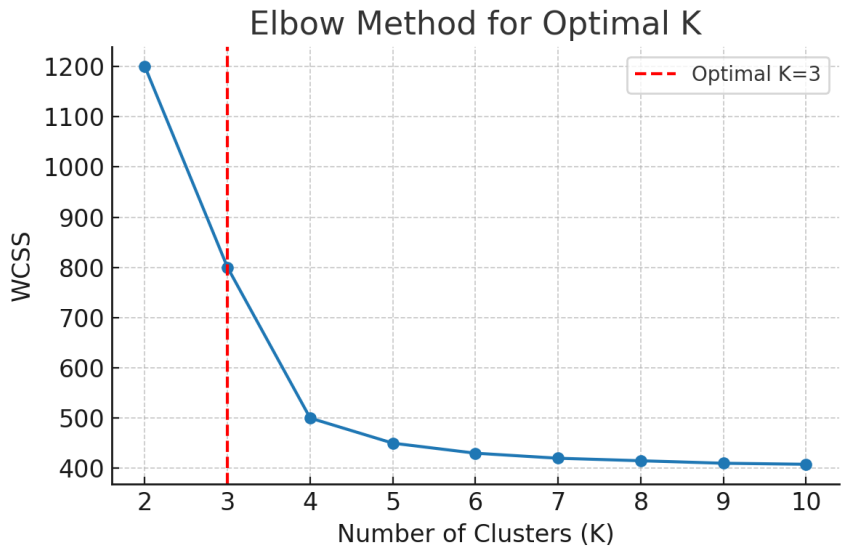
Pada gambar 4 menampilkan hasil klasterisasi dengan algoritma K-Means yang divisualisasikan menggunakan metode Principal Component Analysis (PCA) dalam dua dimensi. Setiap titik merepresentasikan data pasien, sedangkan warna biru (Cluster 0), hijau (Cluster 2), dan oranye (Cluster 1) menunjukkan klaster yang terbentuk. Dari plot terlihat bahwa ketiga klaster memiliki wilayah distribusi yang cukup jelas, meskipun terdapat sedikit tumpang tindih di area perbatasan. Klaster biru cenderung terdistribusi di sisi kiri, klaster hijau berada di bagian tengah, dan klaster oranye mendominasi sisi kanan. Pola ini menunjukkan bahwa data berhasil dipisahkan ke dalam tiga kelompok yang berbeda, sesuai dengan hasil analisis penentuan jumlah klaster optimal sebelumnya (K=3). Visualisasi PCA ini membantu memperkuat interpretasi bahwa pemisahan klaster sudah memadai, meskipun masih terdapat area transisi di mana karakteristik data antar klaster berdekatan.

Evaluasi Hasil Klasterisasi

Silhouette Score yang dihasilkan dari klasterisasi ini adalah sekitar 0.47, menunjukkan pemisahan antar klaster cukup baik, meskipun terdapat beberapa tumpang tindih antar kelompok.Interpretasi WCSS juga menunjukkan penurunan signifikan hingga K = 3, kemudian perlambatan penurunan setelahnya, yang memperkuat pemilihan jumlah klaster optimal tersebut.

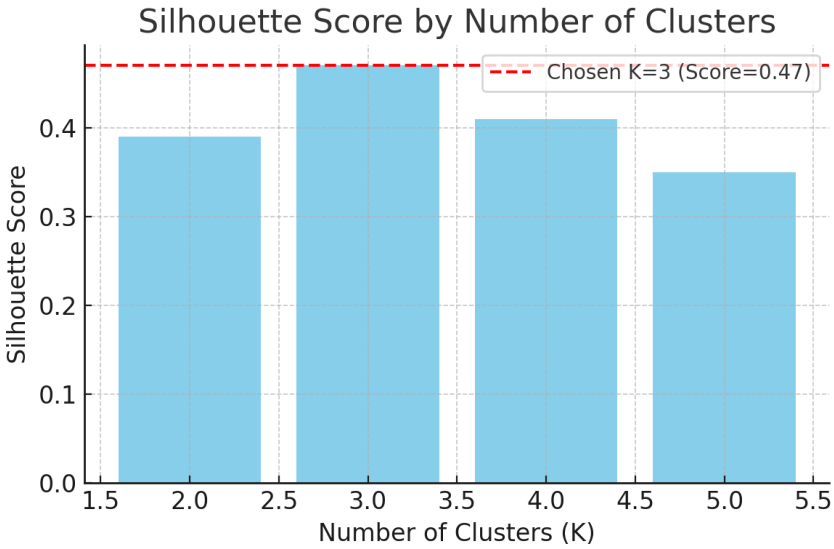
Dari hasil klasterisasi, dapat disimpulkan:

1. Klaster 0 (Pasien Kronis): Usia lebih tua, jumlah diagnosa dan biaya lebih tinggi, dominan menggunakan rawat inap.
2. Klaster 1 (Pasien Rutin): Usia menengah, layanan rawat jalan dominan, kemungkinan pasien kontrol berkala.
3. Klaster 2 (Pasien Umum): Usia lebih muda, biaya dan diagnosa sedikit lebih rendah, juga dominan rawat inap (kemungkinan kasus mendadak atau IGD ringan).



Gambar 5. Elbow method Optimal for K Final

Gambar 5 di atas menunjukkan hasil penerapan metode *Elbow* untuk menentukan jumlah klaster optimal pada algoritma K-Means. Sumbu horizontal menggambarkan jumlah klaster (K) dari 2 hingga 10, sedangkan sumbu vertikal menunjukkan nilai *Within-Cluster Sum of Squares (WCSS)*. Terlihat bahwa nilai WCSS mengalami penurunan tajam dari K=2 ke K=3, kemudian penurunannya mulai melandai pada K berikutnya. Pola ini membentuk sudut siku (elbow) yang jelas pada K=3, sehingga ditetapkan jumlah klaster optimal adalah tiga. Dengan demikian, hasil analisis Elbow Method mendukung pemilihan K=3 sebagai jumlah klaster yang paling efisien, karena menambah jumlah klaster lebih dari tiga tidak lagi memberikan pengurangan signifikan terhadap nilai WCSS.



Gambar 6. Silhouette Score

Gambar 6 Silhouette Score memperlihatkan kualitas hasil klasterisasi untuk berbagai jumlah klaster (K). Sumbu horizontal menunjukkan jumlah klaster yang diuji, sementara sumbu vertikal menunjukkan nilai Silhouette Score rata-rata. Terlihat bahwa nilai tertinggi dicapai pada K=3 dengan skor sekitar 0,47. Nilai ini menunjukkan bahwa klaster yang terbentuk memiliki koherensi internal yang cukup baik serta pemisahan yang memadai antar klaster, meskipun masih terdapat sedikit tumpang tindih pada beberapa data. Jika jumlah klaster ditambah lebih dari tiga, nilai Silhouette Score justru menurun, menandakan pemisahan yang semakin kurang optimal. Dengan demikian, hasil ini memperkuat analisis metode Elbow bahwa K=3 adalah jumlah klaster yang paling sesuai untuk dataset pasien yang digunakan.

Evaluasi hasil klasterisasi dilakukan dengan menggabungkan metode *Elbow*, *Silhouette Score*, visualisasi PCA, serta analisis profil klaster. Berdasarkan grafik Elbow, jumlah klaster optimal ditetapkan pada K=3, karena penurunan nilai *Within-Cluster Sum of Squares (WCSS)* mulai melandai setelah titik tersebut. Hal ini menunjukkan bahwa penggunaan lebih dari tiga klaster tidak lagi memberikan perbaikan signifikan terhadap kualitas klaster.

Hasil ini diperkuat oleh perhitungan *Silhouette Score* yang mencapai nilai rata-rata sebesar 0,47 pada K=3. Skor tersebut mengindikasikan bahwa klaster yang terbentuk memiliki koherensi internal yang cukup baik dan pemisahan antar klaster yang memadai, meskipun masih terdapat beberapa area transisi antar kelompok pasien.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa algoritma K-Means dengan K=3 mampu menghasilkan segmentasi pasien yang cukup baik, valid, dan relevan untuk mendukung strategi pelayanan kesehatan berbasis data.

PEMBAHASAN

Hasil klasterisasi pasien menggunakan algoritma K-Means dengan jumlah klaster optimal sebanyak 3 kelompok memberikan informasi yang sangat berharga dalam konteks manajemen pelayanan rumah sakit. Setiap klaster menunjukkan pola karakteristik pasien yang berbeda, baik dari segi usia, intensitas kunjungan, biaya pengobatan, maupun jenis layanan yang digunakan. Hal ini membuka peluang bagi rumah sakit untuk menerapkan pendekatan yang lebih personalized dan efisien dalam menangani pasien.

Tabel 8 Capaian Pembahasan Hasil Klasterisasi Pasien

Aspek	Klaster 0 (Pasien Kronis)	Klaster 1 (Pasien Rutin)	Klaster 2 (Pasien Umum/Non-Kronis)
Karakteristik Utama	Usia lanjut, banyak diagnosa, biaya tinggi, sering rawat inap	Stabil, kontrol berkala, rawat jalan, biaya sedang	Usia muda, kasus ringan, rawat inap sesekali, sering IGD
Strategi Penanganan	Monitoring penyakit kronis, kunjungan rumah, edukasi gizi, pengobatan jangka panjang	Telemedicine, layanan kontrol rawat jalan terjadwal	Edukasi preventif, gaya hidup sehat, vaksinasi, deteksi dini penyakit
Efisiensi Operasional	Fokuskan alokasi tempat tidur & dokter spesialis	Penjadwalan ulang kontrol rutin, efisiensi jadwal rawat jalan	Penguatan pencegahan & layanan insidental untuk mengurangi beban IGD
Pengembangan Layanan Digital	Sistem peringatan dini, integrasi pemantauan kondisi kronis	Pengingat digital jadwal kontrol, dashboard kunjungan rutin	Kampanye edukatif digital, aplikasi pendaftaran IGD ringan & triase cepat

SIMPULAN

Penelitian ini membuktikan bahwa penerapan algoritma K-Means mampu menghasilkan segmentasi pasien yang relevan dan bermakna berdasarkan kombinasi atribut riwayat kesehatan dan jenis layanan kesehatan yang digunakan. Dengan menggunakan data dari 1.459 pasien yang mencakup informasi demografis, penyakit kronis, intensitas kunjungan, serta jenis layanan kesehatan, model klasterisasi berhasil mengidentifikasi tiga klaster utama pasien, yaitu pasien kronis, pasien rutin, dan pasien umum/non-kronis. Setiap klaster memiliki karakteristik yang khas dan dapat digunakan untuk mendukung pengambilan keputusan di tingkat manajerial rumah sakit, seperti

alokasi sumber daya, pengembangan layanan, hingga desain intervensi yang lebih personal.

Tabel 9 Profil Klaster Pasien

Klaster	Umur Rata – Rata	Rata – Rata Biaya	Rata – Rata Diagnosa	Jenis Layanan Dominan
0	52.4 tahun	Rp 4.974.399	4.17	Rawat Inap
1	50.3 tahun	Rp 5.029.521	4.07	Rawat Jalan
2	49.5 tahun	Rp 4.930.267	3.96	Rawat Inap

Klaster 0 (pasien kronis) cenderung lebih tua, memiliki banyak diagnosa, dan sering menggunakan layanan rawat inap dengan biaya pengobatan yang tinggi, sehingga strategi pelayanan yang disarankan adalah monitoring jangka panjang dan pendekatan homecare. Klaster 1 (pasien rutin) didominasi pasien kontrol berkala yang stabil dengan dominasi layanan rawat jalan, cocok untuk layanan telemedicine dan pengingat digital terjadwal. Klaster 2 (pasien umum) mencakup pasien lebih muda dengan kasus ringan dan rawat inap insidental, cocok untuk edukasi preventif dan layanan triase cepat. Hasil ini menjawab permasalahan utama dalam penelitian, yaitu bagaimana mengelompokkan pasien secara optimal dengan mempertimbangkan berbagai atribut kesehatan dan layanan. Selain itu, penelitian ini memberikan kontribusi penting dalam mendukung pengembangan sistem informasi kesehatan berbasis data dan berpotensi diaplikasikan dalam strategi layanan adaptif rumah sakit di masa depan. Dengan segmentasi yang akurat, rumah sakit dapat merancang kebijakan yang lebih efektif, efisien, dan terfokus pada kebutuhan tiap kelompok pasien secara spesifik.

REFERENCES

[1] R. Purba *et al.*, "Journal of Artificial Intelligence and Engineering Applications Grouping Medical Record Data By Type Diseases With K-Means Algorithm," 2022. [Online]. Available: https://ioinformatic.org/

[2] Y. Syahra, D. Rahman Habibie, M. Nasution, H. N. Nasution, and A. Hadi Nasyuha, "Klasterisasi Data Penanganan dan Pelayanan Kesehatan Masyarakat dengan Algoritma K-Means," *J. Ris. Komputer)*, vol. 9, no. 5, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i5.4888.

[3] D. Novaliendry, T. Wibowo, N. Ardi, T. Evi, and D. Admojo, "Optimizing Patient Medical Records Grouping through Data Mining and K-Means Clustering Algorithm: A Case Study at RSUD Mohammad Natsir Solok," *Int. J. online Biomed. Eng.*, vol. 19, no. 12, pp. 144–155, 2023, doi: 10.3991/ijoe.v19i12.42147.

[4] R. Kaur, P. Kaur, and A. Professor, "K-MEANS CLUSTERING AlGORITHM USING INITIALIZATION AND NORMALIZATION METHODS," 2018. [Online]. Available: www.ijcrt.org

[5] I. D. Borlea, R. E. Precup, and A. B. Borlea, "Improvement of K-means Cluster Quality by Post Processing Resulted Clusters," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 63–70. doi: 10.1016/j.procs.2022.01.009.

[6] M. Bellezza, A. di Palma, and A. Frosini, "Predicting Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using K-Means Clustering on MRI Data," *Inf.*, vol. 15, no. 2, Feb. 2024, doi: 10.3390/info15020096.
[7] X. Chen, "K-Mean Clustering," 2020.
[8] S. S. Momahhed, S. Emamgholipour Sefiddashti, B. Minaei, and Z. Shahali, "K-means clustering of outpatient prescription claims for health insureds in Iran," *BMC Public Health*, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12889-023-15753-1.
[9] A. Simanjuntak and M. S. Hasibuan, "Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram," vol. 11, no. 4, p. 1002, 2023, doi: 10.33394/j.
[10] J. Wala and R. Umar, "Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung," 2024.
[11] S. Faiayaz Waris and S. Koteeswaran Professor, "An Investigation on Disease Diagnosis and Prediction by Using Modified KMean clustering and Combined CNN and ELM Classification Techniques Article History."
[12] P. Supratman, "PENERAPAN METODE K-MEANS UNTUK MENGELOMPOKKAN REKAM MEDIS PASIEN BERDASARKAN DIAGNOSA PENYAKIT GUNA MENENTUKAN DIAGNOSA TERTINGGI PADA SUATU PERIODE (Study Kasus : Klinik Dokter Kita)," 2024.
[13] S. A. Francis, M. Sangeetha, P. Pooranadevi, and A. Sathyakala, "Mathematical Insights and Applications of K-Means Clustering in Diabetes Data Analysis," 2025. [Online]. Available: https://internationalpubls.com
[14] Arshleen, "K-Means Clustering Techniques-A Review." [Online]. Available: https://pramanaresearch.org/
[15] N. El, F. Ravat, and O. Teste, "KD-means: clustering method for massive data based on kd-tree," 2020.
[16] S. Kolay, K. S. Ray, and A. Chand Mondal, "K+ Means : An Enhancement Over K-Means Clustering Algorithm."
[17] B. Suharjo, M. Satria, and Y. Utama, "JISA (Jurnal Informatika dan Sains) K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel," 2021.
[18] H. Dilawati, H. Widianto, and A. Kuswiadji, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering Di Rumah Sakit Widodo Ngawi," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, pp. 139–147, Sep. 2024, doi: 10.37148/bios.v5i2.134.
[19] P. Jat and K. Jain, "A Revised and efficient K-means Clustering Algorithm," *Int. J. Comput. Sci. Eng. Open Access Res. Pap.*, no. 6, 2018, [Online]. Available: www.ijcseonline.org
[20] S. Jain, M. Sharma, and P. Kumar, "Recommendation system for breast cancer treatment using K-means clustering algorithm," *Indian J. Public Heal. Res. Dev.*, vol. 10, no. 4, pp. 202–207, Apr. 2019, doi: 10.5958/0976-5506.2019.00690.9.
[21] Hasudungan, A., Muliono, R., Khairina, N., & Novita, N. (2024). The Impact of k-means on Association Rules Mining Algorithms Performance. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, *5*(2), 640-653.
[22] Muliono, R., & Sembiring, Z. (2019). Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. *CESS (Journal of Computer Engineering, System and Science)*, *4*(2), 2502-2714.