INCODING: Journal of Informatic and Computer Science Engineering
https://journal.mahesacenter.org/index.php/incoding/index || ISSN 2776-432X (online)
5 (2) 2025: 293-304

DOI: 10.34007/incoding.v5i2.983

Analisis Persebaran Penyakit di Wilayah Menggunakan Algoritma K-Means Berbasis Data Kunjungan Fasilitas Kesehatan

Analysis of Disease Distribution in Regions Using the K-Means Algorithm Based on Health Facility Visit Data

Zatin Suhaira & Rizki Muliono*
Program Studi Teknik Informatika, FakultasTeknik, Universitas Medan Area, Indonesia

Diterima: 16 April 2025; Direview: 20 April 2025; Disetujui: 24 April 2025
*Coresponding Email: rizkimuliono@staff.uma.ac.id

Abstrak

Penelitian ini bertujuan untuk menganalisis persebaran penyakit berdasarkan data kunjungan pasien ke berbagai fasilitas kesehatan dengan menggunakan metode *K-Means clustering*. Data penelitian diperoleh secara sekunder dari platform *Kaggle* yaitu '*Healthcare Dataset*' yang berisi informasi pasien, termasuk atribut fasilitas kesehatan, kondisi medis, dan data lainnya. Penentuan jumlah cluster optimal dilakukan dengan menggunakan *Elbow Method*, sedangkan kualitas klasterisasi dievaluasi dengan dua metrik internal, yaitu *Silhouette Score* dan *Davies–Bouldin Index* (DBI). Hasil klasterisasi menghasilkan tiga cluster utama dengan karakteristik berbeda. Cluster pertama didominasi pasien dengan diagnosis arthritis pada kelompok umur 55–59 tahun dengan golongan darah O+. Cluster kedua menampilkan dominasi obesitas pada kelompok umur 35–39 tahun dengan golongan darah AB+, sedangkan cluster ketiga menunjukkan kasus kanker pada kelompok umur 65–69 tahun dengan golongan darah O-. Evaluasi menghasilkan nilai *Silhouette Score* sebesar 0,5349 dan DBI sebesar 0,5830, yang menunjukkan bahwa hasil klasterisasi cukup baik dengan cluster yang kompak dan terpisah. Temuan ini tidak hanya memperlihatkan variasi pola persebaran penyakit di setiap fasilitas kesehatan, tetapi juga dapat dijadikan dasar dalam memetakan distribusi penyakit serta mendukung pengambilan keputusan strategis di bidang kesehatan masyarakat.
Kata Kunci: K-Means; Fasilitas Kesehatan; Rekam Medis; Cluster; Silhouette Score; Davies-Bouldin Index

Abstract

This study aims to analyze the distribution of diseases based on patient visit data to various healthcare facilities using the K-Means clustering method. The research data were obtained secondarily from the Kaggle platform, namely the 'Healthcare Dataset', which contains patient information, including healthcare facility attributes, medical conditions, and other related data. The determination of the optimal number of clusters was carried out using the Elbow Method, while the quality of clustering was evaluated with two internal metrics, namely the Silhouette Score and the Davies–Bouldin Index (DBI). The clustering results produced three main clusters with distinct characteristics. The first cluster was dominated by patients diagnosed with arthritis in the age group of 55–59 years with blood type O+. The second cluster showed a predominance of obesity in the age group of 35–39 years with blood type AB+, while the third cluster indicated cancer cases in the age group of 65–69 years with blood type O-. The evaluation resulted in a Silhouette Score of 0.5349 and a DBI of 0.5830, indicating that the clustering quality is fairly good, with compact and well-separated clusters. These findings not only highlight variations in disease distribution across healthcare facilities but also provide a foundation for mapping disease patterns and supporting strategic decision-making in public health..
Keywords: *K-Means; Health Facilities; Medical Records; Cluster; Silhouette Score; Davies-Bouldin Index*

PENDAHULUAN

Kesehatan memiliki peran yang sangat vital dalam kehidupan manusia dan diakui sebagai salah satu hak asasi yang harus dipenuhi bagi setiap orang tanpa terkecuali [1]. Sebagai bagian dari upaya menjamin hak tersebut, pemantauan terhadap kondisi kesehatan masyarakat secara menyeluruh menjadi sangat penting untuk dilakukan. Berbagai penyakit masih menjadi tantangan serius di Indonesia, dengan pola persebaran yang sering kali dipengaruhi oleh faktor lingkungan dan sosial. Sanitasi yang buruk, akses air bersih yang tidak memadai, serta kebiasaan hidup yang kurang higienis dapat meningkatkan risiko penyebaran penyakit di suatu wilayah. Selain itu, status sosial ekonomi dan tingkat pendidikan masyarakat juga turut berkontribusi terhadap kerentanan suatu daerah terhadap penyakit tertentu [2].

Oleh karena itu, memahami persebaran penyakit secara spasial menjadi langkah penting dalam mendukung upaya pencegahan dan pengendalian yang lebih tepat sasaran di berbagai wilayah. Fasilitas kesehatan merupakan tempat penyelenggaraan pelayanan kesehatan yang meliputi upaya promotif, preventif, kuratif, dan rehabilitatif yang dikelola oleh pemerintah maupun masyarakat. Salah satu komponen penting di dalamnya adalah rekam medis, yaitu dokumen yang memuat informasi lengkap terkait identitas, pemeriksaan, pengobatan, serta tindakan medis yang wajib dicatat dan dijaga kerahasiaannya oleh tenaga medis [3]. Pengisian rekam medis harus dilakukan secara lengkap dan tepat waktu karena kualitas pelayanan di fasilitas kesehatan tercermin dari kelengkapan dokumen tersebut. Rekam medis memiliki nilai penting, baik bagi fasilitas kesehatan, tenaga medis, maupun pasien yang menerima layanan [4].

Data kunjungan dan catatan kesehatan masyarakat yang terdokumentasi di fasilitas pelayanan kesehatan dapat menjadi sumber informasi yang berharga, karena mencatat jenis penyakit, waktu kunjungan, serta lokasi asal pasien secara sistematis. Dengan adanya data yang terstruktur tersebut, analisis persebaran penyakit di tingkat wilayah dapat dilakukan secara lebih akurat, sehingga mempermudah perencanaan intervensi dan peningkatan mutu layanan kesehatan masyarakat [5]. Pemanfaatan data kunjungan pada fasilitas kesehatan menjadi salah satu pendekatan penting dalam mengkaji pola persebaran penyakit berdasarkan karakteristik wilayah tertentu.

Penelitian sebelumnya telah menunjukkan bahwa analisis persebaran penyakit berbasis spasial dapat membantu dalam mengidentifikasi daerah rawan secara lebih

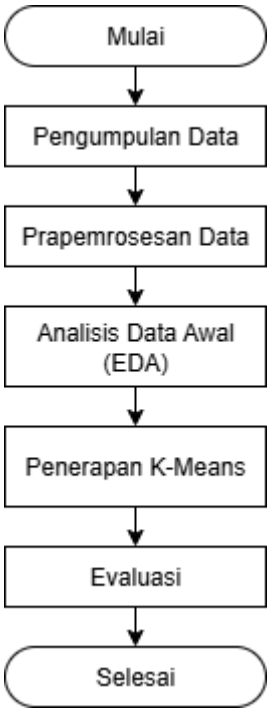
sistematis. Salah satu studi menerapkan teknik data mining untuk memetakan wilayah yang terdampak penyakit demam berdarah dengue (DBD), dengan memanfaatkan data kasus dari beberapa desa di Kecamatan Setia Janji. Melalui proses klasterisasi, wilayah tersebut berhasil dikategorikan ke dalam tiga kelompok berdasarkan tingkat penyebarannya: tinggi, sedang, dan rendah. Algoritma *K-Means* digunakan dalam proses ini dan menghasilkan performa optimal dengan nilai *Davies-Bouldin Index* sebesar 1.044. Hasil pemetaan yang diperoleh dapat digunakan oleh pemerintah untuk mengambil tindakan secara cepat dan tepat dalam menangani wilayah yang memiliki kerentanan tinggi terhadap penyebaran penyakit DBD di Kecamatan Setia Janji [6].

Temuan dari penelitian tersebut menjadi dasar penting bagi penelitian ini, yang akan menerapkan pendekatan serupa untuk menganalisis persebaran penyakit secara umum, dengan memanfaatkan data kunjungan fasilitas kesehatan sebagai indikator utama dalam pengelompokan wilayah. Dengan menerapkan algoritma *K-Means*, penelitian ini diharapkan dapat memberikan gambaran spasial mengenai daerah-daerah yang memiliki tingkat kunjungan tinggi akibat penyakit tertentu, sehingga dapat digunakan sebagai dasar perencanaan intervensi kesehatan yang lebih tepat sasaran.

METODE PENELITIAN

Penelitian ini dilakukan melalui beberapa tahapan sistematis untuk menghasilkan klasterisasi yang akurat terhadap data persebaran penyakit berdasarkan kunjungan fasilitas kesehatan. Tahap pertama adalah pengumpulan data sekunder yang diperoleh dari platform *Kaggle*, yang memuat informasi pasien beserta atribut terkait seperti umur, jenis kelamin, jenis penyakit (*Medical Condition*), rumah sakit (*Hospital*), jenis layanan (*Admission Type*), dan tanggal kunjungan. Selanjutnya, dilakukan prapemrosesan data yang mencakup pembersihan data, penanganan nilai hilang, pengelompokan berdasarkan lokasi rumah sakit sebagai representasi wilayah, dan transformasi data ke dalam format numerik yang sesuai untuk analisis. Setelah itu, dilakukan *Exploratory Data Analysis* (EDA) untuk memahami distribusi penyakit per wilayah, tren kunjungan pasien, serta hubungan antar variabel. Tahap implementasi mencakup penerapan algoritma *K-Means* terhadap data yang telah diproses, dengan penentuan jumlah cluster optimal menggunakan metode *Davis Bouldin Index* (DBI). Terakhir, hasil klasterisasi dievaluasi menggunakan visualisasi peta serta analisis karakteristik tiap cluster untuk menilai sejauh mana model mampu

membedakan kelompok wilayah berdasarkan pola persebaran penyakit yang teridentifikasi. Gambar 1 di bawah ini menunjukkan tahapan dalam penelitian



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh secara sekunder dari platform Kaggle dengan judul '*Healthcare Dataset*'. Dataset ini memuat informasi terkait pasien yang menjalani perawatan di berbagai fasilitas kesehatan, yang mencakup atribut seperti nama pasien, usia, jenis kelamin, golongan darah, kondisi medis, tanggal masuk, dokter penanggung jawab, rumah sakit, penyedia asuransi, jumlah tagihan, nomor kamar, jenis masuk (*admission type*), tanggal keluar, obat yang diberikan, serta hasil tes medis. Karena dataset ini tidak menyertakan informasi koordinat geografis pasien, pemetaan persebaran penyakit pada difokuskan pada lokasi rumah sakit atau fasilitas kesehatan, bukan pada pembagian wilayah administrasi.

Prapemrosesan Data

Tahap Pra-pemrosesan data adalah proses dimana menyeleksidata yang tidak sesuai dan mengubah data tersebut menjadi bentuk yang lebih mudah diproses oleh sistem [7]. Tahapan pra-pemrosesan data dilakukan untuk menjamin kualitas dan kesiapan data kesehatan sebelum digunakan dalam proses klasterisasi persebaran penyakit [8].

EDA

Exploratory Data Analysis (EDA) merupakan proses analisis dan visualisasi data yang bertujuan untuk memperoleh pemahaman yang lebih mendalam mengenai karakteristik dan wawasan yang terkandung dalam data [9]. Proses *exploratory data analysis* (EDA) umumnya mencakup pemeriksaan manual, penyajian data secara visual, serta penerapan berbagai metode statistik untuk memahami karakteristik dan pola pada data [10]. Berdasarkan hasil EDA, tahap selanjutnya adalah melakukan pengelompokan data menggunakan metode *K-Means* untuk menemukan struktur dan pola yang lebih jelas pada dataset.

Penerapan K-Means

Metode *K-Means clustering* merupakan salah satu teknik dalam data mining, sehingga dapat disimpulkan bahwa semakin besar jumlah data yang diolah, maka kualitas hasil yang diperoleh juga akan semakin optimal. Penggunaan data yang lebih banyak dalam pengujian *K-Means* dapat menghasilkan prediksi yang lebih akurat [11]. Algoritma *K-Means* merupakan metode klasterisasi yang bersifat iteratif. Algoritma ini menggunakan jarak sebagai standar pengukuran untuk membagi data menjadi K kelompok, menghitung rata-rata jarak setiap kelompok, lalu menentukan titik pusat (*centroid*) awal. Setiap kelompok selanjutnya direpresentasikan oleh titik pusat tersebut [12]. Pada algoritma *K-Means*, setiap data ditempatkan ke dalam suatu cluster tertentu dan dapat berpindah ke cluster lain pada iterasi berikutnya. Metode ini bersifat non-hierarkis, di mana tahap awalnya menentukan pusat cluster dari sebagian data dalam populasi. Pemilihan pusat cluster awal dilakukan secara acak dari kumpulan data yang tersedia [9].

Setiap cluster direpresentasikan oleh titik pusat (centroid). Tujuan dari proses ini adalah membentuk kelompok-kelompok terpisah dari n titik data $\{x_1, x_2, \ldots, x_n\}$ menjadi k (< n) himpunan $\{S_1, S_2, \ldots, S_k\}$ dengan meminimalkan nilai rata-rata total, termasuk jarak kuadrat antara setiap titik dengan centroid-nya. Dengan demikian, tujuan optimasinya adalah untuk menemukan konfigurasi cluster yang menghasilkan nilai minimum tersebut [12].

$$\underset{S}{argmin} \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Dimana:

S = Himpunan cluster $\{S_1, S_2, \ldots, S_k\}$
n = Jumlah total data
k = Jumlah cluster
x = Titik data
μ_i = Centroid cluster
$\|x - \mu_i\|$ = Jarak euclidean kuadrat antara titik data x dan μ_i

Untuk menentukan jumlah cluster k yang optimal, digunakan metode *Elbow Method*. *Elbow Method* merupakan pendekatan evaluasi hasil klasterisasi dengan melihat hubungan antara jumlah cluster dan nilai *Within-Cluster Sum of Squares* (WCSS). WCSS menggambarkan tingkat kohesi, yaitu seberapa dekat data dalam satu cluster terhadap *centroid*-nya. Semakin kecil nilai WCSS, semakin baik kohesi antar data dalam cluster. Namun, penambahan jumlah cluster secara terus-menerus akan selalu menurunkan nilai WCSS. Oleh karena itu, *Elbow Method* mencari titik siku (*elbow point*) pada grafik jumlah cluster terhadap WCSS, di mana penurunan WCSS mulai melambat. Dalam penerapannya, metode ini dilakukan dengan menghitung selisih kuadrat (SSE) pada berbagai nilai k yang diuji, misalnya dari 1 hingga 10 cluster, sehingga titik siku yang terbentuk dapat menunjukkan jumlah cluster yang paling optimal [13][14]. Formula untuk mencari nilai *Elbow Method* adalah sebagai berikut.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Dimana:
C_i = Cluster ke-i
μ_i = Centroid cluster ke-i

Evaluasi

Evaluasi hasil klasterisasi dilakukan untuk menilai kualitas dan validitas pemisahan antar cluster yang dihasilkan oleh algoritma *K-Means*. Penelitian ini menggunakan dua metrik evaluasi internal, yaitu *Silhouette Score* dan *Davies–Bouldin Index* (DBI). DBI adalah metode evaluasi hasil klasterisasi yang mempertimbangkan nilai kohesi dan separasi. Kohesi mengacu pada tingkat kedekatan data terhadap centroid cluster yang ditempatinya, sedangkan separasi menggambarkan jarak antar centroid dari masing-masing cluster [15]. Pengklasteran dianggap baik apabila menghasilkan nilai DBI serendah mungkin, karena hal ini menunjukkan bahwa setiap cluster memiliki tingkat kekompakan tinggi dan terpisah dengan jelas dari cluster lainnya [16]. Formula yang digunakan untuk mencari DBI adalah sebagai berikut.

$$DBI = \frac{1}{K}\sum_{i=1}^{k} max_{i \neq j}(R_{ij})$$

Dimana:
K = Cluster
R_{ij} = Rasio antar cluster i dan j
Max = Rasio antar cluster yang terbesar

Silhouette Score mengukur tingkat kesesuaian suatu titik data dengan klasternya dibandingkan dengan cluster lain, sedangkan DBI menilai rasio antara kedekatan titik-titik data dalam cluster (kohesi) dan keterpisahan antar cluster (separasi). Silhouette Score merupakan metrik yang digunakan untuk menilai kualitas hasil klasterisasi, dengan asumsi bahwa cluster yang baik memiliki bentuk yang kompak serta terpisah dengan jelas satu sama lain [17].

Silhouette Score digunakan untuk menilai seberapa baik sebuah data cocok dengan klasternya sendiri dibandingkan dengan cluster lain. Skor ini berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menandakan bahwa data tersebut sangat sesuai dengan klasternya, sedangkan nilai yang mendekati -1 mengindikasikan kemungkinan data tersebut salah ditempatkan dalam cluster. Berikut adalah rumus Silhouette Score untuk sebuah titik i:

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

dengan:
a(i) = rata-rata jarak antara i dengan semua titik lain dalam klaster yang sama (intra-cluster distance)
b(i) = jarak rata-rata terkecil dari titik i ke semua titik dalam klaster lain yang paling dekat (narest-cluster distance)

HASIL DAN PEMBAHASAN

Pemrosesan Data

Pemrosesan data dalam penelitian ini diawali dengan pemeriksaan nilai kosong (*missing value*) untuk memastikan kualitas data, diikuti dengan pemilihan atribut yang relevan dengan analisis persebaran penyakit berdasarkan fasilitas kesehatan serta penghapusan atribut yang tidak berkontribusi signifikan. Data yang digunakan mencakup rentang waktu 8 Mei 2019 hingga 7 Mei 2024. Atribut kategorikal, seperti jenis penyakit dan kategori fasilitas kesehatan, dikonversi ke bentuk numerik menggunakan teknik label

encoding agar dapat diproses oleh algoritma *K-Means*. Selanjutnya dilakukan eksplorasi data (EDA) melalui analisis statistik deskriptif pada variabel kategori untuk melihat jumlah data, variasi nilai, serta frekuensi kemunculan, dan analisis pada variabel usia untuk memperoleh informasi mengenai rentang, rata-rata, serta distribusi usia pasien. Hasil EDA kemudian divisualisasikan melalui diagram distribusi usia, diagnosis, dan golongan darah, sehingga memberikan gambaran awal mengenai profil demografis dan kondisi kesehatan pasien yang menjadi dasar penting dalam tahap klasterisasi.

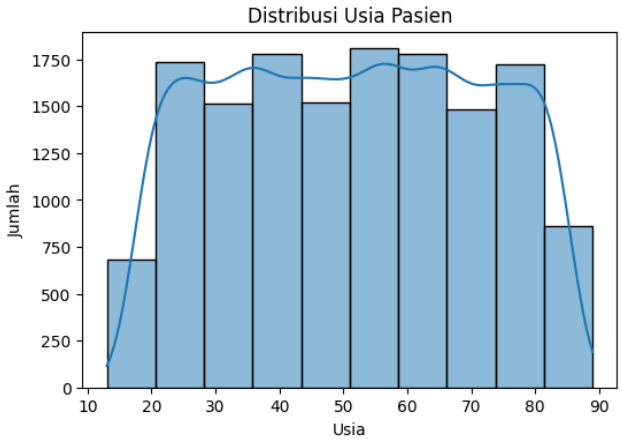
Tabel 1. Statistik Deskriptif Variabel Kategorikal

	Jumlah	Unik	Top	Freq
Hospital (Faskes)	14.880	12.017	Smith Inc	16
Medical Condition (Diagnosa)	14.880	6	Arthritis	2.517
Blood Type (Gol. Darah)	14.880	8	AB+	1.909

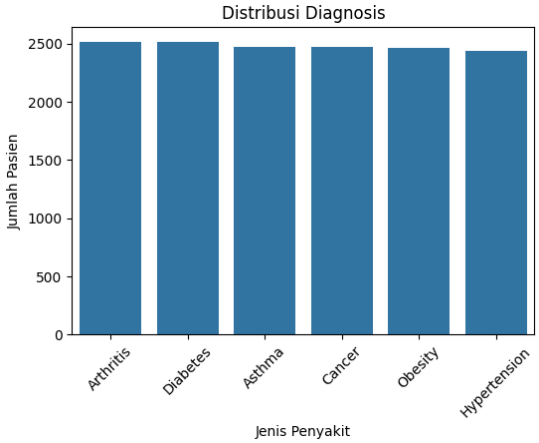
Tabel 2. Statistik Deskriptif Variabel Usia Pasien

	Jumlah	Mean	Std	Min	25%	50%	75%	Max
Umur	14.880	51	20	13	35	52	68	89

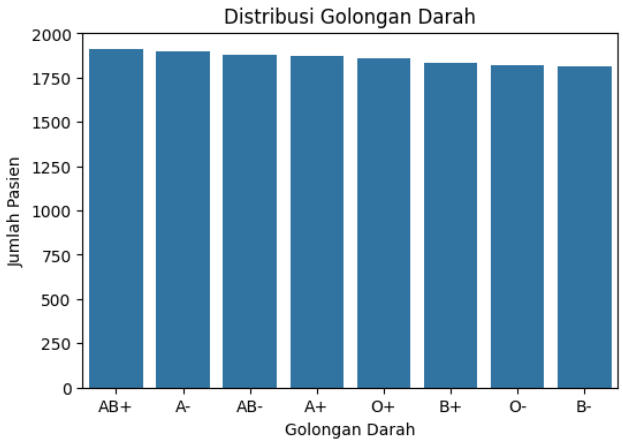
Berdasarkan ringkasan statistik di atas, terlihat variasi karakteristik pasien baik dari segi kategori maupun usia, yang selanjutnya divisualisasikan melalui plot distribusi untuk memperoleh gambaran yang lebih jelas.



(a)



(b)



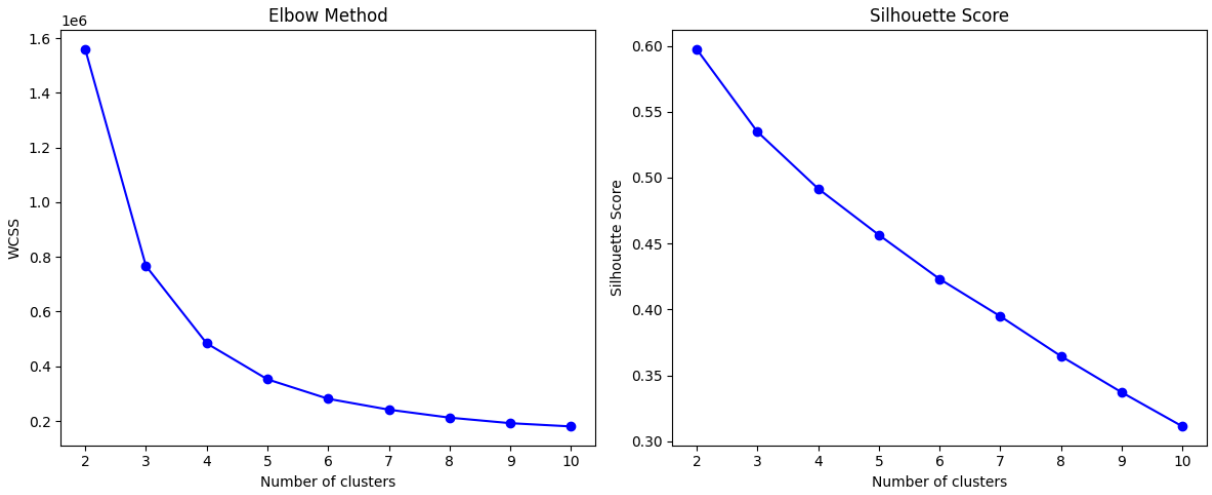
(c)

Gambar 2. Distribusi Karakteristik Pasien Berdasarkan (a) Usia, (b) Diagnosa, dan (c) Golongan Darah

Berdasarkan hasil eksplorasi data yang dilakukan, diperoleh gambaran awal mengenai karakteristik pasien yang meliputi persebaran usia, jenis diagnosa, golongan darah, serta fasilitas kesehatan yang dikunjungi. Informasi ini menjadi dasar penting dalam proses pemodelan dan analisis lebih lanjut pada penelitian ini. Selanjutnya, metode K-Means digunakan untuk mengelompokkan data pasien berdasarkan karakteristik tersebut guna mengidentifikasi pola persebaran penyakit secara spasial. Pendekatan ini diharapkan dapat memberikan wawasan yang lebih mendalam dalam pengelompokan wilayah berdasarkan data kunjungan fasilitas kesehatan.

Penerapan K-Means

Sebelum menerapkan algoritma *K-Means*, terlebih dahulu dilakukan penentuan jumlah cluster optimal menggunakan *metode Elbow Method*. Metode ini mengevaluasi hubungan antara jumlah cluster dan nilai *Sum of Squared Errors* (SSE) atau *Within-Cluster Sum of Squares* (WCSS). Seiring bertambahnya jumlah cluster, nilai SSE/WCSS akan menurun, namun pada titik tertentu penurunannya mulai melambat. Titik siku (*elbow point*) pada grafik jumlah cluster terhadap SSE/WCSS inilah yang dipilih sebagai jumlah cluster optimal, karena dianggap memberikan keseimbangan antara kekompakan cluster dan kompleksitas model. Gambar berikut memperlihatkan perubahan nilai SSE pada berbagai kandidat jumlah cluster, sehingga dapat diidentifikasi titik siku sebagai jumlah cluster optimal.



Gambar 3. Evaluasi jumlah cluster optimal menggunakan Elbow Method dan Silhouette Score

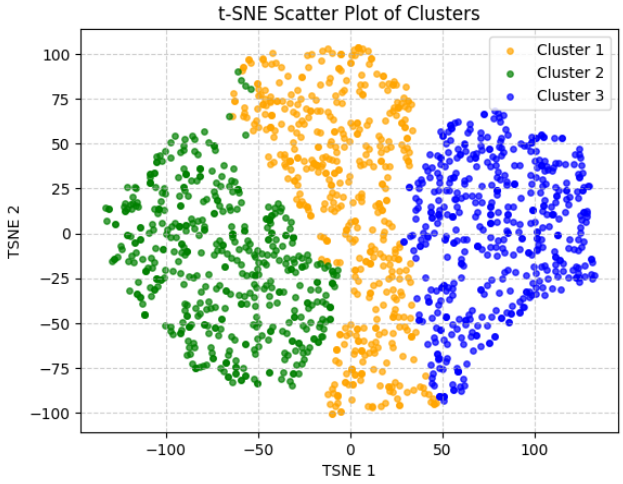
Gambar 3 menunjukkan hubungan antara jumlah cluster (k) dengan nilai *Within-Cluster Sum of Squares* (WCSS) pada *Elbow Method* serta *Silhouette score* untuk setiap kandidat cluster. Titik siku pada grafik *Elbow* dan nilai *Silhouette* yang relatif stabil mengindikasikan bahwa jumlah cluster optimal berada pada k=3. Oleh karena itu, jumlah cluster k=3 dipilih sebagai konfigurasi optimal untuk proses klasterisasi.

Setelah diperoleh jumlah cluster optimal, proses klasterisasi dilakukan menggunakan nilai *k* tersebut. Hasil klasterisasi disajikan pada Tabel 3.

Tabel 3. Karakteristik utama setiap cluster berdasarkan karakteristik pasien yang paling dominan.

Cluster	Umur	Golongan Darah	Diagnosis
1	55–59 tahun	O+	Arthritis
2	35–39 tahun	AB+	Obesity
3	65-69 tahun	O-	Cancer

Tabel 3 menyajikan karakteristik utama dari masing-masing cluster berdasarkan fasilitas kesehatan, kelompok umur, golongan darah, dan diagnosis yang paling dominan. Hasil analisis menunjukkan bahwa Cluster 1 didominasi oleh pasien dengan kelompok umur 55–59 tahun, golongan darah O+, dan diagnosis arthritis. Cluster 2 memiliki karakteristik pasien dengan kelompok umur 35–39 tahun, golongan darah AB+, dan diagnosis obesitas. Sementara itu, Cluster 3 ditandai oleh pasien dengan kelompok umur 65–69 tahun, golongan darah O-, dan diagnosis kanker. Visualisasi menggunakan t-SNE ditampilkan untuk memperjelas pemisahan antar cluster.



Gambar 4. Visualisasi hasil klasterisasi menggunakan t-SNE yang menunjukkan distribusi data pada tiga cluster.

Evaluasi

Evaluasi kualitas hasil klasterisasi dilakukan dengan bantuan program *Python* menggunakan dua metrik internal, yaitu *Silhouette Score* dan *Davies-Bouldin Index* (DBI). Hasil perhitungan disajikan pada Tabel berikut.

Table 4. Hasil Evaluasi

Metrik	Nilai
Silhouette Score	0,5349
Davies-Bouldin Index (DBI)	0,5830

SIMPULAN

Berdasarkan hasil analisis, klasterisasi berhasil mengelompokan fasilitas kesehatan berdasarkan karakteristik pasien, seperti kelompok umur, golongan darah, dan diagnosis dominan. Hasil klasterisasi menunjukkan adanya variasi pola persebaran penyakit di tiap fasilitas kesehatan, yang merepresentasikan perbedaan karakteristik pasien dan jenis penyakit yang lebih menonjol pada masing-masing cluster. Evaluasi menggunakan *Silhouette Score* sebesar 0,5349 dan *Davies–Bouldin Index* sebesar 0,5830 mengindikasikan bahwa kualitas klasterisasi berada pada tingkat cukup baik, dengan cluster yang relatif kompak dan terpisah. Temuan ini memperlihatkan bahwa algoritma *K-Means* dapat dimanfaatkan untuk mengidentifikasi pola distribusi penyakit secara lebih terstruktur.

Kedepannya, pendekatan ini dapat dikembangkan dengan menambahkan variabel spasial atau data longitudinal, sehingga analisis mampu memantau dinamika persebaran penyakit dari waktu ke waktu dan diterapkan pada skala wilayah yang lebih luas guna mendukung pengambilan keputusan di bidang kesehatan masyarakat.

DAFTAR PUSTAKA

[1] G. B. Mentari and Susilawati, "Faktor-faktor Yang Mempengaruhi Akses Pelayanan Kesehatan di Indonesia," J. Heal. Sains, vol. 3, no. 8.5.2017, pp. 2003–2005, 2022.

[2] A. S. Mentari and B. Besral, "Analisis Faktor Higiene Sebagai Sumber Penularan Hepatitis a Di Indonesia: Literature Review," J. Cahaya Mandalika ISSN 2721-4796, pp. 2293–2301, 2024, [Online]. Available: https://ojs.cahayamandalika.com/index.php/jcm/article/view/3209

[3] S. Sofia, E. T. Ardianto, N. Muna, and S. Sabran, "Analisis Aspek Keamanan Informasi Data Pasien Pada Penerapan RME di Fasilitas Kesehatan," J. Rekam Med. Manaj. Inf. Kesehat., vol. 1, no. 2, pp. 94–103, 2022, doi: 10.47134/rmik.v1i2.29.

[4] R. Marbun, R. Ariyanti, and V. Dea, "Peningkatan Pengetahuan Masyarakat Terkait Pentingnya Rekam Medis Bagi Pasien Di Fasilitas Pelayanan Kesehatan," SELAPARANG J. Pengabdi. Masy. Berkemajuan, vol. 5, no. 1, p. 163, 2021, doi: 10.31764/jpmb.v5i1.6427.

[5] W. T. Ina, Y. Mesakh, and S. I. Pella, "Klusterisasi Penyakit Endemis Pada Kecamatan Sabu Barat, Kabupaten Sabu Raijua Menggunakan Algoritma K-Means," J. Media Elektro, vol. XI, no. 1, pp. 39–44, 2022, doi: 10.35508/jme.v11i1.6508.

[6] M. A. Sembiring, "Penerapan Metode Algoritma K-Means Clustering Untuk Pemetaan Penyebaran Penyakit Demam Berdarah Dengue (DBD)," J. Sci. Soc. Res., vol. 4, no. 3, p. 336, 2021, doi: 10.54314/jssr.v4i3.712.

[7] F. Muhammad, N. M. Maghfur, and A. Voutama, "Sentiment Analysis Dataset on COVID-19 Variant News," Systematics, vol. 4, no. 1, pp. 382–391, 2022.

[8] A. Amato and V. Di Lecce, "Data preprocessing impact on machine learning algorithm performance," Open Comput. Sci., vol. 13, no. 1, p. 20220278, 2023.

[9] M. Dzaki Salman, N. Rizki Pratama, M. A. Nakhlah Farid, A. Agung Setiawan, F. Zalianti, and I. Bil Huda, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison of K-Means and K-Medoids Clustering Algorithm Performance in Grouping Schools in Riau Province Based on Availability of Facilities and Infrastructure," vol. 5, no. July, pp. 797–806, 2025.

[10] F. C. Oettl, J. F. Oeding, R. Feldt, C. Ley, M. T. Hirschmann, and K. Samuelsson, "The artificial intelligence advantage: Supercharging exploratory data analysis," Knee Surgery, Sport. Traumatol. Arthrosc., vol. 32, no. 11, pp. 3039–3042, 2024, doi: 10.1002/ksa.12389.

[11] R. Muliono and Z. Sembiring, "Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen," J. Comput. Eng. Syst. Sci., vol. 4, no. 2, pp. 2502–714, 2019.

[12] B. Chong, "K-means clustering algorithm: a brief review," Acad. J. Comput. Inf. Sci., vol. 4, no. 5, pp. 37–40, 2021, doi: 10.25236/ajcis.2021.040506.

[13] P. M. Hasugian, B. Sinaga, J. Manurung, and S. A. Al Hashim, "Best Cluster Optimization with Combination of K-Means Algorithm And Elbow Method Towards Rice Production Status Determination," Int. J. Artif. Intell. Res., vol. 5, no. 1, pp. 102–110, 2021, doi: 10.29099/ijair.v6i1.232.

[14] V. A. Permadi, S. P. Tahalea, and R. P. Agusdin, "K-Means and Elbow Method for Cluster Analysis of Elementary School Data," Prog. Pendidik., vol. 4, no. 1, pp. 50–57, 2023, doi: 10.29303/prospek.v4i1.328.

[15] I. Irwan, W. Sanusi, A. S. Anwar, and A. Rahman, "The Implementation of Spatial Model with K-Means Clustering Method to Cluster Flood Affected Areas in Bone Regency," ARRUS J. Soc. Sci. Humanit., vol. 3, no. 2, pp. 186–195, 2023, doi: 10.35877/soshum1771.

[16] M. Wahyudi, S. Solikhun, and L. Pujiastuti, "Komparasi K-Means Clustering dan K-Medoids Clustering dalam Mengelompokkan Produksi Susu Segar di Indonesia Berdasarkan Nilai DBI," J. Bumigora Inf. Technol., vol. 4, no. 2, pp. 243–254, 2022, doi: 10.30812/bite.v4i2.2104.

[17] G. Vardakas, I. Papakostas, and A. Likas, "Deep Clustering Using the Soft Silhouette Score: Towards Compact and Well-Separated Clusters," 2024, [Online]. Available: http://arxiv.org/abs/2402.00608