Computer Science (CO-SCIENCE)
Vol. 5 No. 2 Juli 2025
Akreditasi Sinta 4 Nomor SK : 230/E/KPT/2022

Studi Komparatif Algoritma K-Means dan K-Medoids untuk Segmentasi Informasi Kesehatan

Muhammad Dwi Ananda[1], Karenina Nurmelita Malik[2], Anis Fitri Nur Masruriyah[3*], Mardiah[4]

[1,2,3,4]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jakarta
Jl. R.S Fatmawati No. 1, Cilandak, Jakarta Selatan 12450, Indonesia

e-mail: [1]2210511089@mahasiswa.upnvj.ac.id, [2]2210511119@mahasiswa.upnvj.ac.id,
[3*]masruriyah@upnvj.ac.id, [4]mardiah@upnvj.ac.id

(*) Corresponding Author

Artikel Info : Diterima : 20-06-2025 | Direvisi : 20-07-2025 | Disetujui : 22-07-2025

Abstrak - Dalam analisis data medis agar mendukung keputusan klinis, diperlukan segmentasi informasi kesehatan yang memegang peranan krusial. Penelitian ini menyajikan analisis komparatif algoritma K-Means dan K-Medoids dalam pengelompokkan data *Medical Examination*. Evaluasi ini dilakukan dengan menggunakan dua pendekatan internal utama, yaitu Silhouette Score dan Davies-Bouldin Index dalam mengukur kualitas pemisahan serta kohesi antar klaster. Eksperimen ini melibatkan berbagai variasi jumlah klaster guna menentukan konfigurasi optimal tiap algoritma. Hasil penelitian menunjukkan bahwa K-Means memberikan performa yang representatif dan lebih stabil terhadap kompleksitas data, dibandingkan dengan algoritma K-Medoids yang hanya optimal dalam jumlah klaster kecil. Analisis statistik menggunakan *one way* ANOVA diterapkan untuk menguji signifikansi perbedaan performa antar algoritma berdasarkan rata-rata nilai Silhouette Score dengan menunjukkan nilai F sebesar 4.8594 dengan P-value sebesar 0.0447. Hal ini mengindikasikan bahwa perbedaan performa kedua algoritma signifikan secara statistik pada tarif signifikansi 5%. Penelitian ini menegaskan dalam algoritma K-Means untuk segmentasi data kesehatan dengan distribusi yang beragam, serta diharapkan mampu menjadi landasan bagi pengembangan sistem klasifikasi data kesehatan yang lebih efisien di masa mendatang.

Kata Kunci : Segmentasi Informasi, Clustering, K-Means, K-Medoids, Data Kesehatan

Abstracts *- In analyzing medical data to support clinical decisions, segmentation of health information plays a crucial role. This study presents a comparative analysis of K-Means and K-Medoids algorithms in clustering Medical Examination data. This evaluation is conducted using two main internal approaches, namely Silhouette Score and Davies-Bouldin Index in measuring the quality of separation as well as cohesion between clusters. The experiment involved varying the number of clusters to determine the optimal configuration of each algorithm. The results show that K-Means provides representative performance and is more stable against data complexity, compared to the K-Medoids algorithm which is only optimal in a small number of clusters. Statistical analysis using one-way ANOVA was applied to test the significance of performance differences between algorithms based on the average Silhouette Score value, yielding an F-value of 4.8594 with a P-value of 0.0447. This indicates that the performance difference between the two algorithms is statistically significant at 5% significance rate. This research confirms the K-Means algorithm for segmenting health data with diverse distributions and is expected to serve as a foundation for the development of more efficient health data classification systems in the future.*

Keywords : Information Segmentation, Clustering, K-Means, K-Medoids, Health Data

PENDAHULUAN

Kemajuan teknologi informasi dan melimpahnya data dalam era Big Data menunjukkan pentingnya analisis data kesehatan dalam pengambilan keputusan yang didasarkan pada bukti di sektor layanan kesehatan. Volume data kesehatan yang besar dan kompleks menimbulkan permasalahan ketika informasi yang bermakna akan diekstraksi dan diolah lebih lanjut. *Clustering* merupakan salah satu metode dalam data mining yang berfungsi sebagai pengelompokkan data ke dalam beberapa kelompok berdasarkan kesamaan karakteristik tertentu tanpa menggunakan label yang telah ditentukan sebelumnya (Hendrastuty, 2024). Teknik ini memungkinkan identifikasi struktur alamiah dan pola tersembunyi dari kumpulan data kesehatan dengan kompleksitas dan dimensi tinggi.

Para ahli dalam bidang kesehatan dapat mengungkap hubungan dan asosiasi yang sebelumnya tidak terdeteksi dalam data pasien dengan menggunakan teknik *clustering*. Dengan demikian, teknik ini dapat membantu pemahaman terhadap pola penyakit, efektivitas terapi, serta mengidentifikasi faktor risiko kesehatan.

Sistem layanan kesehatan di lapangan menghasilkan berbagai jenis data seperti rekam medis, hasil laboratorium, dan data sensor pasien dengan jumlah yang terus bertambah (Purba et al., 2023). Akan tetapi, data yang terlalu berlimpah seringkali tidak disertai dengan pemrosesan atau analisis yang sesuai, sehingga potensi informasi yang efektif tidak dapat digunakan secara optimal. Hal ini menyebabkan rendahnya efektivitas pengambilan keputusan klinis, sulit dalam mendeteksi dini mengenai risiko penyakit, hingga efisiensi layanan menjadi rendah secara keseluruhan.

Perkembangan teknologi di bidang informatika berhasil membuka peluang besar dalam pengelolaan data kesehatan, terkhusus dengan pendekatan *data mining* yang mampu mengeksplorasi pengetahuan tersembunyi dari dataset besar (Sari et al., 2024). Pada ranah ini, ada berbagai teknik yang telah dikembangkan untuk mendukung pengambilan keputusan data, salah satunya adalah *clustering*. Di antara beberapa algoritma *clustering*, K-Means dan K-Medoids merupakan dua alternatif yang populer, hal ini disebabkan kemampuan untuk memisahkan data menjadi beberapa kelompok yang bermakna berdasarkan kesamaan karakteristik. Akan tetapi, kedua algoritma ini memiliki karakteristik dan keterbatasan yang berbeda satu sama lain, sehingga akan mempengaruhi kualitas hasil segmentasi pada konteks data kesehatan. K-Means merupakan algoritma yang bekerja dengan memisahkan data sesuai dengan jumlah klaster (K) yang telah ditentukan, kemudian diawali dengan seleksi acak pada K titik pusat, lalu mengalokasikan data berdasarkan kedekatannya dengan pusat (Fadil & Fatah, 2025). Di sisi lain, K-Medoids adalah algoritma yang sering direferensikan sebagai Partitioning Around Medoids (PAM) dalam menerapkan pendekatan yang berbeda (Ningrum et al., 2021). Meskipun keduanya terkategorisasi sebagai algoritma partitioning, K-Medoids menggunakan titik data aktual (medoid) berasal dari himpunan data untuk representasi klaster. Hal ini berbeda dengan algoritma K-Means yang menggunakan nilai rerata (centroid). Dengan demikian, penting dilakukan studi komparatif dalam menentukan algoritma yang paling sesuai pada konteks pengelompokan data kesehatan.

Penelitian yang dilakukan oleh (Permatasari et al., 2024) menunjukkan bahwa algoritma K-Means dan K-Medoids efektif dalam mengelompokkan data pada balita stunting berdasarkan atribut-atribut, seperti tinggi badan atau berat badan. Hasil pada algoritma K-Means menunjukkan performa baik dengan nilai DBI bernilai 0.0005. (Syamfithriani et al., 2023) dalam penelitiannya menerapkan pendekatan dengan dua algoritma dalam pemetaan wilayah prioritas penanganan penyakit diare pada balita di Kabupaten Kuningan. Hasil dalam penelitian ini adalah algoritma K-Means lebih unggul dibandingkan algoritma K-Medoids pada evaluasi DBI. Tidak hanya itu, K-Means dinilai lebih efektif ketika mengidentifikasi daerah prioritas untuk penanganan penyakit.

Penelitian yang dilakukan oleh (Meiriza et al., 2023) berfokus dalam pengelompokan program BPJS non-upah menggunakan algoritma k-Means dan K-Medoids. Hasil dalam penelitian ini adalah K-Medoids menghasilkan klaster yang lebih stabil dengan nilai DBI dibandingkan dengan algoritma K-Means. Penelitian sebelumnya yang dilakukan oleh (Christnatalis - et al., 2023) menggunakan kedua algoritma dalam mengukur kepuasan pelayanan rumah sakit. Kuesioner menampilkan kategori “sangat puas” dengan nilai DBI K-Medoids = 2.17, sedangkan nilai K-Means = 2.70. Hal ini menunjukkan jika K-Medoids cenderung memberikan pengelompokkan yang lebih akurat.

Pengelompokkan wilayah berbasis data merupakan salah satu pendekatan yang penting untuk penanganan isu kesehatan masyarakat, seperti penelitian oleh (Fira et al., 2021) dengan data COVID-19 dari 34 provinsi di Indonesia yang berlangsung dari tahun 2019 hingga 2021 untuk mengelompokkan wilayah berdasarkan tingkat kasus. Dari perbandingan algoritma K-Means dan K-Medoids menunjukkan jika K-Medoids memiliki nilai Silhouette Coefficient yang lebih tinggi dengan nilai 0.347, sedangkan untuk K-Means mendapatkan nilai 0.207. Menurut penelitian (Nirwana et al., 2022) dimana mereka mengkaji zonasi wilayah COVID-19 di Sumatera Selatan dan mendapatkan jika K-Means memberikan hasil yang lebih optimal dengan DBI sebesar 0.078 dibandingkan K-Medoids dengan DBI 0.250. Kedua penelitian sama-sama menampilkan keunggulan K-Means pada efisiensi dan keakuratan dalam pemetaan wilayah kesehatan.

Penelitian oleh (Momahhed et al., 2023) dalam mengelompokkan 21.776.350 kali resep rawat jalan dengan algoritma K-Means berdasarkan data demografis dan penggunaan obat. Hasil segmentasi mendapatkan adanya tiga klaster risiko yang diimplementasikan dalam penetapan premi mengelola risiko kerugian. Sementara itu, (Leis et al., 2023) menggunakan algoritma K-Medoids dalam pengelompokkan pasien influenza rawat inap berdasarkan data laboratorium dan parameter fisiologis. Hasil yang didapatkan adalah klaster dengan kadar glukosa tinggi menampilkan jika masa rawat inap lebih panjang dan risiko terhadap pemakaian ventilator.

Menurut (Safitri, 2024) dalam penelitian membandingkan dua algoritma *clustering* untuk mengelompokkan jumlah tenaga kesehatan rumah sakit di Bojonegoro, algoritma K-Means terbukti lebih efektif dibandingkan algoritma K-Medoids Hasil pengelompokan K-Means terbagi menjadi empat klaster berdasarkan nilai jarak rata-rata dalam klaster dengan merepresentasi tingkat dari rendah ke tinggi. Penelitian oleh (Utomo, 2021) dalam analisis penyebaran COVID-19 di Indonesia dari data Kementerian Kesehatan menunjukkan algoritma K-Means

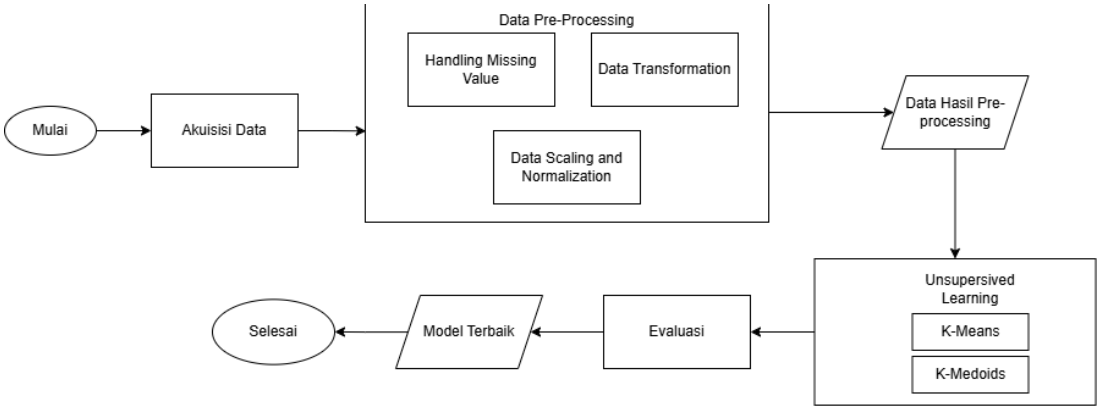
memiliki nilai Davies-Bouldin terendah yaitu 0.064 pada k = 5, sementara itu dalam algoritma K-Medoids memiliki nilai Davies-Bouldin 0.411 pada k = 2.

Penelitian ini bertujuan untuk membandingkan efektivitas antara algoritma K-Means dan K-Medoids dalam mengklasifikasikan data kesehatan. Dilatarbelakangi oleh kebutuhan yang semakin meningkat terhadap implementasi data mining dalam manajemen informasi kesehatan, khususnya untuk mendukung sistem pengambilan keputusan klinis. Penelitian ini menganalisis berbagai parameter kesehatan seperti tinggi badan, berat badan, tekanan darah sistolik dan diastolik, serta indikator lainnya yang diperoleh dari dataset pemeriksaan medis. Penilaian efektivitas kedua algoritma dilakukan melalui dua metrik validitas internal utama, yaitu Silhouette Score dan Davis-Bouldin Index yang disertai dengan analisis terhadap efisiensi waktu pemrosesan. Meskipun algoritma K-Means dan K-Medoids telah banyak diterapkan dalam penelitian sebelumnya, penelitian ini memberikan kontribusi tambahan melalui dengan mengevaluasi performa pada berbagai jumlah klaster, memanfaatkan visualisasi berbasis *Principal Component Analysis* (PCA), serta penggunaan kombinasi metrik evaluasi secara bersamaan untuk memperoleh analisis yang lebih menyeluruh. Diharapkan hasil penelitian ini dapat memberikan rekomendasi algoritma yang lebih unggul dalam klasifikasi data kesehatan, serta memberikan fondasi yang kuat bagi penerapan sistem berbasis bukti dalam dunia medis.

METODE PENELITIAN

1. Alur Penelitian

Melalui pendekatan sistematis, penelitian ini dilakukan mengikuti rangkaian tahapan yang dimulai dari akuisisi data hingga evaluasi model yang diilustrasikan pada Gambar 1.



Sumber: Hasil Penelitian (2025)

Gambar 1. Alur Penelitian

Penelitian diawali dengan mengakuisisi data dari yang bersumber dari Kaggle, kemudian data akan dilakukan tahap preprocessing untuk menangani data yang hilang (*missing value*), melakukan normalisasi, dan transformasi data. Selanjutnya, data akan dilakukan proses *unsupervised learning* dengan algoritma K-Means dan K-Medoids. Model yang dihasilkan akan dievaluasi untuk menentukan hasil terbaik metrik tertentu untuk menentukan hasil optimal. Model terbaik kemudian dimanfaatkan dalam interpretasi dan penarikan kesimpulan.

2. Akuisisi Dataset

Akuisisi data merupakan tahap awal dalam metodologi data mining, dimana proses ini melibatkan pengumpulan dataset primer yang diperlukan dalam proyek data tersebut. Dataset awal ini akan menjalani transformasi menjadi dataset final sesudah melalui serangkaian proses analisis dan preparasi data. Dataset final kemudian akan diaplikasikan pada fase pemodelan agar menghasilkan model yang diharapkan (Wahyudi et al., 2022).

Tabel 1. Deskripsi Dataset

No	Atribut	Penjelasan
1	*Age*	Parameter krusial dalam studi yang menunjukkan umur individu dalam satuan hari
2	*Sex*	Mengidentifikasi jenis kelamin, memiliki implikasi dalam perbedaan struktural tubuh dan predisposisi penyakit

No	Atribut	Penjelasan
3	*Height*	Pengukuran tinggi dalam sentimeter
4	*Weight*	Massa tubuh dalam kilogram
5	*Ap_Hi*	Parameter tekanan darah sistolik yang menggambarkan tekanan darah saat memompa ke pembuluh arteri
6	*Ap_Lo*	Parameter tekanan darah diastolik yang menunjukkan tekanan pada arteri saat jantung istirahat
7	*Cholesterol*	Indikator kadar kolesterol dalam aliran darah
8	*Gluc*	Nilai kadar gula darah yang berisiko diabetes
9	*Smoke*	Indikator perilaku merokok
10	*Alco*	Parameter konsumsi alkohol yang dapat mempengaruhi tekanan darah, fungsi hati, dan meningkatkan risiko penyakit kardiovaskular
11	*Active*	Indikator tingkat aktivitas fisik
12	*Cardio*	Aktivitas yang meningkatkan detak jantung dan pernapasan

Sumber: (Hasan, 2025)

3. Preprocessing Data

Proses dalam *preprocessing* dimulai dengan tahapan Pembersihan Data (*Data Cleaning*) yang memiliki tujuan utama untuk menghilangkan data *noise* serta mengatasi data yang kurang konsisten agar kualitas data tetap optimal dalam proses analisis (Romli, 2021). Selain itu, tahap ini juga mencakup proses integrasi data, yaitu penyatuan informasi yang berasal dari berbagai sistem atau sumber yang berbeda-beda menjadi satu kesatuan dataset yang konsisten dan utuh. Proses ini sangat penting untuk memastikan bahwa data yang berasal dari klien, vendor, ataupun sumber daring seperti situs web dapat diolah secara terpusat guna mendukung analisis data secara menyeluruh (Joshi & Patel, 2021). Dalam penelitian ini, proses pembersihan mencakup penghilangan nilai-nilai yang tidak relevan, penghapusan data duplikat dan kolom identifikasi yang tidak memiliki kontribusi dalam analisis.

Tahap selanjutnya adalah melakukan Transformasi terhadap Fitur Kategorikal menggunakan metode *One-Hot Encoding*. Teknik ini bertujuan agar nilai yang telah dikonversi kategorikal menjadi representasi numerik yang dapat diproses oleh algoritma *machine learning*. Masing-masing kategori dalam atribut akan diubah menjadi kolom biner yang berisi nilai 0 atau 1 (Setiawan et al., 2025). Beberapa atribut yang ditransformasikan adalah *cardio, alco, sex, active, cholesterol, smoke,* dan *gluc*. Seperti contoh, atribut *gluc* memiliki tiga kategori yang akan ditransformasi menjadi tiga kolom baru, yaitu *gluc_1, gluc_2,* dan *gluc_3*.

Berikutnya adalah melakukan tahap Normalisasi Data, proses untuk menyamakan skala semua fitur numerik sehingga algoritma tidak bias dalam salah satu fitur dengan rentang nilai yang besar. Pada tahap ini menggunakan teknik *StandardScaler*, di mana teknik ini akan menghasilkan data dengan nilai tengah 0 dan standar deviasi 1. Normalisasi ini penting untuk pengolahan data secara sederhana dan mempercepat proses pembelajaran mesin (Permana & Salisah, 2022).

Dalam tahapan *preprocessing* juga dilakukan Deteksi dan Penanganan Outlier, yaitu data yang berada jauh dari distribusi umum dalam populasi. Keberadaan *outlier* memberikan dampak negatif karena dapat mempengaruhi distribusi data menjadi tidak normal, menyebabkan bias dalam taksiran parameter, serta mempengaruhi validitas pengujian (Razaki et al., 2024). Dengan demikian, visualisasi seperti *boxplot* digunakan untuk mendeteksi *outlier* yang kemudian dapat ditangani secara kontekstual.

Dalam tahap terakhir dilakukan Visualisasi Korelasi antar fitur menggunakan *heatmap*. Visualisasi ini dapat membantu untuk melihat sejauh mana korelasi antar variabel dan mendukung proses pemilihan fitur yang relevan. Pemahaman mengenai pola dalam hubungan antar fitur sangat penting sebelum dilakukannya proses pemodelan (Samosir et al., 2021).

4. Unsupervised Learning

Unsupervised learning merupakan salah satu pendekatan *machine learning* yang digunakan tanpa memerlukan data berlabel, sehingga model secara otomatis mengenali pola dalam data tanpa informasi kelas tertentu (Nurhalizah et al., 2024). Pada penelitian ini dilakukan *preprocessing* sebagai tahapan awal untuk meningkatkan kualitas data, termasuk penanganan data hilang (*missing value*) menggunakan metode imputasi, yaitu metode dengan mengisi nilai modus untuk data kategorikal dan nilai rata-rata untuk data numerik, serta standarisasi agar seluruh fitur berada dalam skala yang sama dan tidak mendominasi proses klasterisasi. Proses berikutnya adalah klasterisasi menggunakan K-Means dan K-Medoids dengan variasi jumlah klaster dari K = 2 hingga K = 10 untuk menentukan jumlah klaster yang paling optimal berdasarkan evaluasi metrik internal.

5. Evaluasi

Hasil klasterisasi dalam penelitian ini menggunakan metrik internal, yaitu Silhouette Score dan Davies-Bouldin Index (DBI). Silhouette Score digunakan untuk menilai konsistensi objek dalam nilai yang lebih tinggi mengindikasikan pemisahan klaster yang lebih baik. DBI mengevaluasi rasio jarak antar klaster terhadap jarak dalam klaster, di mana nilai yang lebih rendah menunjukkan kualitas klasterisasi yang lebih optimal. Untuk memastikan bahwa perbedaan hasil metrik pada berbagai jumlah klaster bersifat signifikan secara statistik, maka diterapkan *Analysis of Variance* (ANOVA) untuk memverifikasi bahwa variasi hasil bukan acak, akan tetapi perbedaan yang signifikan secara statistik. ANOVA merupakan metode uji parametrik yang digunakan dalam mengidentifikasi perbedaan rata-rata antara dua atau lebih kelompok data dengan dengan menganalisis dan membandingkan nilai variansi masing-masing kelompok (Arif et al., 2023).

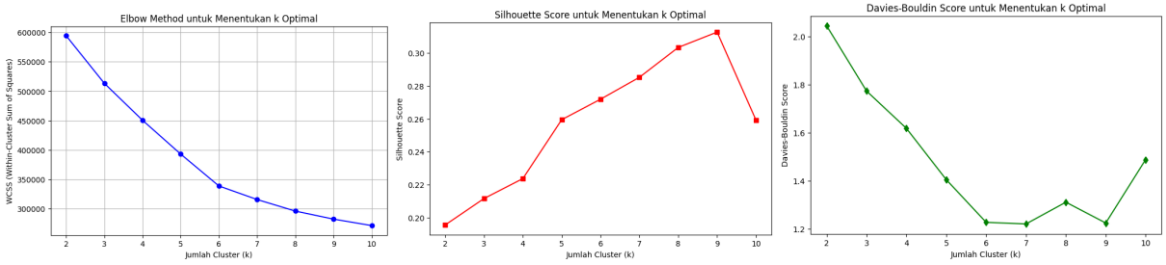
HASIL DAN PEMBAHASAN

Dalam dataset yang digunakan, beberapa langkah pra-pemrosesan data dilakukan untuk memastikan data siap digunakan dalam algoritma *clustering*. Salah satu tahap awal yang penting adalah transformasi kolom usia (*age*), di mana semula usia pasien dicatat pada satuan hari. Representasi dalam satuan hari memang lebih presisi, akan tetapi pada konteks analisis kesehatan, penggunaan usia dalam tahun jauh lebih intuitif dan bermakna. Sebagai contoh, usia 18.393 hari dikonversi menjadi 50 tahun dengan membagi angka hari tersebut dengan 365.25 (mengikuti standar perhitungan tahun kabisat). Konversi ini bertujuan agar menghasilkan distribusi umur yang lebih mudah diinterpretasikan dan lebih relevan dalam mendeteksi pola risiko kesehatan.

Langkah berikutnya adalah melakukan transformasi atribut-atribut kategorikal, seperti *gluc, alco, active, cardio, sex, smoke,* dan *cholesterol* dengan menggunakan metode one-hot encoding. Metode ini dilakukan agar menghindari bias jarak dalam algoritma *clustering* yang berbasis numerik, hal ini dikarenakan nilai 1, 2, dan 3 bukan representasi urutan, melainkan kategori. Sebagai contoh, nilai 1 pada kolom *gluc* berarti kadar glukosa normal, sedangkan nilai nilai 3 adalah kadar tingginya, akan tetapi bukan berarti 3 adalah tiga kali lebih besar dari 1.

Sebelum melakukan ke proses *clustering*, seluruh data numerik akan dilakukan normalisasi dengan menggunakan StandardScaler. Proses ini akan menyamakan rata-rata menjadi 0 dan standar deviasi menjadi 1. Normalisasi ini penting dilakukan agar tidak ada fitur yang mendominasi perhitungan jarak dalam proses pembentukan klaster.

Tahapan awal dalam eksplorasi (EDA) dilakukan dengan menampilkan visualisasi distribusi data dan memahami karakteristik pada setiap fitur. Hasil visualisasi dalam boxplot dan histogram akan menampilkan adanya outlier dalam data, sedangkan untuk heatmap memvisualisasi korelasi antara atribut-atribut. Dengan demikian, seluruh fitur dinilai relevan untuk dianalisis lebih lanjut.



Sumber: Hasil Penelitian (2025)

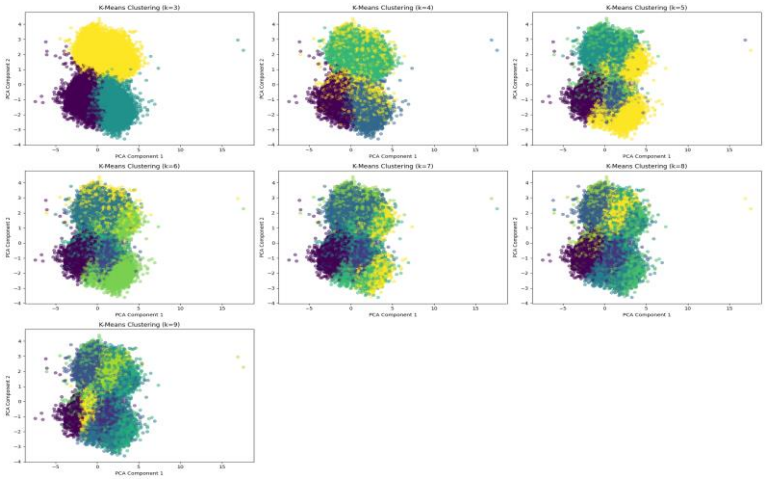
(a) (b) (c)

Gambar 2. Penentuan Jumlah Klaster dalam K-Means

Pada Gambar 2 ditampilkan hasil eksperimen penentuan jumlah klaster optimal dengan tiga pendekatan, (a) Elbow Method, (b) Silhouette Score, dan (c) Davies-Bouldin Index. Meskipun masing-masing pendekatan memberikan indikasi yang berbeda, akan tetapi pola umum dari grafik tersebut memberikan landasan untuk menentukan rentang jumlah klaster yang berhubungan dengan proses *clustering* berikutnya.

Grafik Elbow (a) menunjukkan penurunan nilai Within-Cluster Sum of Squares (WCSS) yang signifikan hingga klaster ke-6, setelahnya penurunan mulai melandai. Peristiwa ini merupakan indikasi keberadaan titik siku (*elbow point*) di sekitar k = 6, dimana sering digunakan sebagai acuan dalam metode ini. Pada grafik Silhouette Score (b), terlihat jika nilai skor mengalami peningkatan seiring bertambahnya jumlah klaster, dimana puncak dalam grafik ini adalah pada nilai tertinggi k = 8. Berdasarkan skor ini menunjukkan jika konfigurasi dengan delapan klaster memberikan keterikatan dalam klaster dan pemisahan antar klaster yang paling ideal berdasarkan metrik.

Sementara itu, grafik Davies-Bouldin Index (c) menunjukkan adanya penurunan nilai DBI secara signifikan, hingga pada nilai k = 6 dan k =7 terjadi kenaikan pada jumlah klaster yang lebih besar. Dikarenakan nilai DBI yang lebih rendah menampilkan kualitas klaster yang lebih baik (lebih padat dan terpisah), sehingga jumlah klaster 6 dan 7 menjadi kandidat kuat berdasarkan metrik ini. Berdasarkan ketiga pendekatan ini, maka jumlah klaster yang akan dieksplorasi lebih lanjut pada rentang k = 3 hingga k = 9. Hal ini bertujuan untuk menangkap keseimbangan antara separabilitas, kompleksitas, dan keterikatan klaster. Hasil evaluasi lanjutan terhadap konfigurasi jumlah klaster dalam rentang tersebut akan dijadikan acuan pada penentuan klaster akhir dengan algoritma K-Means sebelum dilakukan perbandingan dengan algoritma K-Medoids.



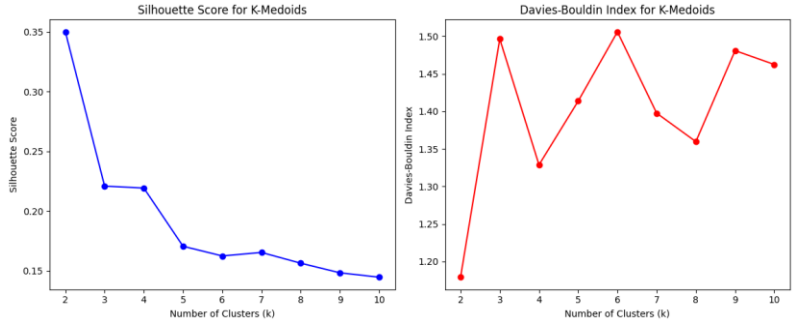
Sumber: Hasil Penelitian (2025)

Gambar 3. Visualisasi K-Means Clustering

Visualisasi hasil *clustering* algoritma K-Means dalam data yang telah melalui reduksi dimensi PCA ditampilkan pada Gambar 3. Terdapat tujuh subplot yang ditampilkan, dimana masing-masing merepresentasikan konfigurasi jumlah klaster dari k = 3 hingga k = 9. Setiap titik-titik ini akan mewakili satu sampel data dan warna yang berbeda menandakan klasifikasi klaster.

Seiring dengan jumlah klaster yang meningkat, pola distribusi terlihat semakin terpisah dan konsisten. Pada konfigurasi k = 3 dan k = 4, terdapat tumpang tindih area klaster secara signifikan. Hal ini menunjukkan apabila pemisahan antar klaster belum optimal. Sementara itu, pada k = 5 hingga k = 7, terlihat peningkatan kualitas pemisahan antar klaster dengan terbentuknya beberapa zona padat yang lebih jelas secara visual. Observasi ini konsisten dengan hasil evaluasi sebelumnya di mana metrik Silhouette Score dan Davies-Bouldin Index menunjukkan performa superior pada kisaran jumlah klaster tersebut.

Visualisasi untuk konfigurasi k = 8 menampilkan distribusi yang lebih tersebar dengan beberapa area klaster yang mulai bersinggungan kembali, meskipun tetap mempertahankan struktur fundamental dari konfigurasi sebelumnya. Pada k = 9, teridentifikasi ada beberapa klaster kecil yang terisolasi, akan tetapi dengan peningkatan tumpang tindih yang berpotensi mengakibatkan terjadinya over-segmentasi data. Melalui visualisasi ini, dapat ditarik kesimpulan bahwa jumlah klaster optimal berada pada rentang 6 hingga 8 dalam merepresentasikan struktur data berdasarkan tampilan spasial dari dua komponen utama PCA. Hasil ini memperkuat analisis sebelumnya pada proses penentuan jumlah klaster optimal dengan algoritma K-Means.



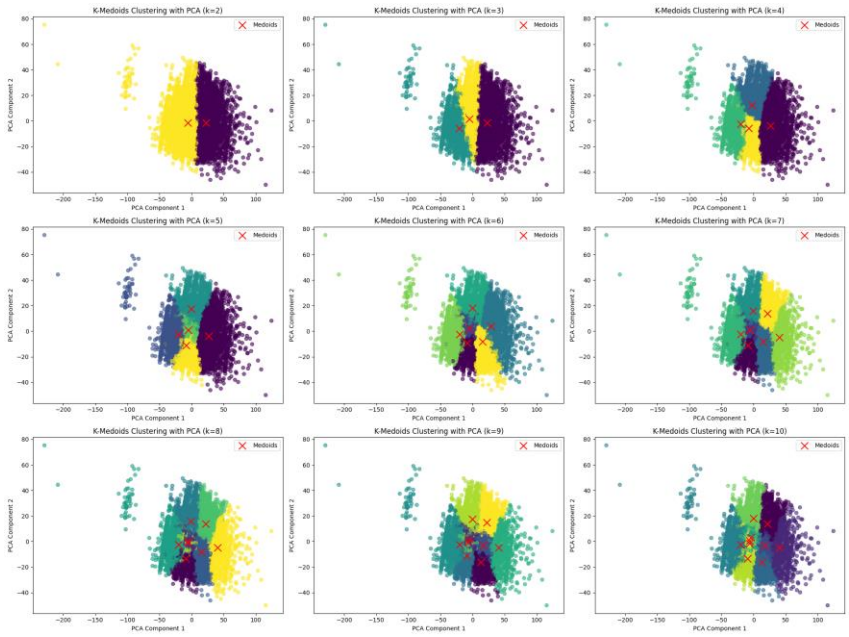
Sumber: Hasil Penelitian (2025)

(a) (b)

Gambar 4. Penentuan Jumlah Klaster dalam K-Medoids

Penentuan jumlah klaster yang ideal dalam algoritma K-Medoids dilakukan dengan dua pendekatan, yaitu Silhouette Score dan Davies-Bouldin Index (DBI). Berdasarkan grafik Silhouette Score (a), nilai tertinggi diperoleh saat k = 2, di mana menunjukkan apabila kualitas pemisahan antar klaster adalah paling paling efektif diantara yang lain. Hal ini terbukti dengan nilai setelahnya yaitu k = 3 hingga k = 10 mengalami penurunan secara konsisten, menandakan bahwa penambahan jumlah klaster kurang meningkatkan kualitas segmentasi dalam Silhouette Score.

Berbanding terbalik dengan pendekatan Davies-Bouldin Index yang menunjukkan nilai terendah pada k = 2, di mana mengindikasikan bahwa klaster yang terbentuk memiliki jarak antar pusat klaster yang baik dengan sebaran internal yang kecil. Nilai DBI mengalami perubahan yang tidak menentu saat k bertambah, namun tidak menunjukkan adanya perbaikan yang signifikan. Berdasarkan kedua metrik ini, nilai k = 2 dipertimbangkan sebagai jumlah klaster yang paling optimal dalam algoritma K-Medoids.



Sumber: Hasil Penelitian (2025)

Gambar 5. Visualisasi K-Medoids Clustering

Visualisasi hasil *clustering* menggunakan algoritma K-Medoids ditampilkan pada Gambar 5 dengan variasi jumlah klaster dari k = 2 hingga k = 10. Metode *Principal Component Analysis* (PCA) digunakan dalam memproyeksikan visual data ke dalam dua dimensi untuk memudahkan pengamatan pola secara visual. Pada konfigurasi k = 2 dan k = 3 menunjukkan jika data terlihat terbagi dalam klaster yang memiliki pemisahan yang cukup jelas.

Klaster yang terbentuk menampilkan distribusi yang padat dan konsisten, di mana sesuai dengan hasil evaluasi yang telah dilakukan sebelumnya. Akan tetapi, ketika nilai k meningkat, distribusi antar klaster mulai menunjukkan overlap yang signifikan tanpa adanya batas yang tegas. Hal ini mengindikasikan adanya pembentukan klaster yang semakin kurang optimal dan segmentasi data menjadi kurang efektif. Berdasarkan hasil visualisasi ini, bahwa nilai k =2 atau k =3 merupakan pilihan optimal dalam menghasilkan struktur klaster yang terpisah dengan baik dan mudah diinterpretasikan dalam penerapan metode K-Medoids.

Tabel 2. Perbandingan Evaluasi K-Means dan K-Medoids

Jumlah Klaster (k)	K-Means Silhouette	K-Means DBI	K-Medoids Silhouette	K-Medoids DBI
2	-	-	0.3500	1.1792
3	0.2116	1.7734	0.2208	1.4965
4	0.2274	1.5503	0.2192	1.3288
5	0.2348	1.4585	0.1704	1.4138

6	0.2529	1.5183	0.1623	1.5059
7	0.2894	1.3128	0.1652	1.3976
8	0.3034	1.3103	0.1562	1.3597
9	0.2502	1.4132	0.1482	1.4810
10	-	-	0.1145	1.4625

Sumber: Hasil Penelitian (2025)

Berdasarkan hasil yang diperoleh dari masing-masing algoritma yang telah dijalankan dan ditampilkan dalam Tabel 2, terlihat perbedaan signifikan dalam kualitas klaster yang dihasilkan oleh masing-masing algoritma. Pada algoritma K-Means, nilai tertinggi dari metode Silhouette Score tercapai dengan klaster k = 9 sebesar 0.3127, hal ini menunjukkan jika pemisahan antar klaster relatif efektif. Tidak hanya itu, nilai Davies-Bouldin Index 9DBI) terendah ditemukan pada jumlah klaster k = 7 dengan nilai 1.2196, hal ini menandakan jika klaster pada jumlah tersebut paling padat dan terpisah secara detail pada konteks algoritma K-Means.

Berbeda dengan algoritma K-Medoids, di mana hasil menunjukkan bahwa performa terbaik terletak dalam jumlah k = 2 dengan nilai Silhouette Score sebesar 0.3500 dan DBI terendah sebesar 1.1792. Akan tetapi, performa algoritma K-Medoids cenderung menurun seiring bertambahnya jumlah klaster, yang menyebabkan dari Silhouette Score dan DBI performanya juga berkurang dalam membagi data menjadi lebih banyak klaster.

Secara keseluruhan, meskipun algoritma K-Medoids terlihat unggul dalam jumlah klaster k = 2, konfigurasi tersebut terlalu sederhana dalam kebutuhan segmentasi informasi kesehatan yang kompleks. Berbanding terbalik dengan algoritma K-Means, di mana algoritma ini dapat menghasilkan kualitas klaster yang lebih baik dan stabil dengan jumlah klaster lebih besar, terkhusus pada k = 8 dan k = 9. Dengan demikian, K-Means dipilih sebagai algoritma yang sesuai pada konteks penelitian ini dikarenakan memberikan keseimbangan antar hubungan dan pemisahan klaster yang lebih efektif.

Perbedaan performa antara algoritma K-Means dan K-Medoids dalam segmentasi data dapat diamati tidak hanya dari nilai evaluasi numerik seperti Silhouette Score dan DBI, tetapi juga dari sudut pandang statistik. Dalam hal ini, penerapan *one-way* ANOVA dilakukan untuk menguji terhadap rata-rata Silhouette Score dari K-Means dan K-Medoids secara statistik. Berdasarkan hasil analisis, diperoleh nilai F sebesar 4.8594 dengan P-value 0.0447. Hal ini menunjukkan bahwa perbedaan performa antara kedua algoritma tidak hanya terlihat secara visual dari nilai rata-rata, tetapi juga signifikan secara statistik pada taraf signifikansi 5%. Temuan ini memperkuat bahwa pemilihan algoritma klasterisasi memiliki dampak nyata terhadap hasil segmentasi data.

Perbedaan performa antara K-Means dan K-Medoids berkaitan erat dengan mekanisme pemilihan pusat klaster. K-Means menggunakan centroid yang dipengaruhi oleh distribusi nilai ekstrem, sehingga lebih sensitif terhadap *outlier*.Akibatnya, K-Means cenderung memiliki variabilitas skor DBI yang tinggi, terutama pada jumlah klaster yang terus bertambah. Sebaliknya, K-Medoids yang memilih titik aktual (medoid) menunjukkan stabilitas lebih tinggi saat jumlah klaster kecil, namun performa K-Medoids menurun dikarenakan ketidakmampuannya menangkap struktur data kompleks saat klaster bertambah.

Meskipun K-Means menunjukkan performa unggul dalam banyak metrik, algoritma ini memiliki keterbatasan utama berupa sensitivitas terhadap outlier serta kecenderungan menghasilkan klaster berbentuk bola (*spherical*). Hal ini dapat mengurangi akurasi segmentasi pada distribusi data yang tidak homogen. Sementara itu, K-Medoids lebih tahan terhadap *outlier* karena menggunakan titik aktual sebagai medoid, namun menjadi kurang efisien saat jumlah klaster besar karena kompleksitas komputasi dan kesulitan dalam menangkap variasi distribusi data yang luas.

KESIMPULAN

Penelitian ini bertujuan untuk membandingkan algoritma K-Means dan K-Medoids pada proses segmentasi informasi kesehatan berdasarkan data *Medical Examination*. Berdasarkan hasil evaluasi metrik dan pengujian statistik, K-Means secara konsisten menghasilkan klaster berkualitas tinggi, terutama pada jumlah klaster yang lebih besar. Pengujian *one-way* ANOVA pada Silhouette Score menghasilkan nilai F = 4.8594 dengan P-value = 0.0447, yang artinya perbedaan performa antara K-Means dan K-Medoids signifikan secara statistik pada taraf signifikansi 5%. Meskipun K-Medoids menunjukkan keunggulan pada konfigurasi dua klaster dan ketahanan terhadap outlier, algoritma ini kurang optimal dalam menangkap struktur data yang kompleks untuk segmentasi data kesehatan yang bervariasi. Dengan demikian, K-Means lebih direkomendasikan untuk segmentasi data kesehatan dalam penelitian dikarenakan kemampuannya menyesuaikan dengan kompleksitas distribusi data.

REFERENSI

Arif, Alfarez, D. A., & Ramadhan, M. R. (2023). Anova dan Tukey HSD Perbandingan Produksi Padi Antara Tiga Kabupaten di Provinsi Jambi. Multi Proximity: Jurnal Statistika, 2(1), Article 1. https://doi.org/10.22437/multiproximity.v2i1.25908

Christnatalis -, Claudyo, E., Lucky -, Manullang, H. K., & Zebua, A. I. (2023). ANALISIS PELAYANAN RUMAH SAKIT UMUM DENGAN PERBANDINGAN ANTARA METODE ALGORITMA KMEANS, DAN K-MEDOIDS CLUSTERING. JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP), 6(2), Article 2. https://doi.org/10.34012/jutikomp.v6i2.4145

Fadil, A., & Fatah, Z. (2025). ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN SISWA UNGGULAN BERDASARKAN HASIL UJIAN DI SEKOLAH. Jurnal Riset Sistem Informasi, 2(1), Article 1. https://doi.org/10.69714/p26gcf27

Fira, A., Rozikin, C., & Garno, G. (2021). Komparasi Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Penyebaran Covid-19 di Indonesia. Journal of Applied Informatics and Computing, 5(2), Article 2. https://doi.org/10.30871/jaic.v5i2.3286

Hasan, S. (2025). Medical Examination Dataset. https://www.kaggle.com/datasets/jazidesigns/medical-examination-dataset

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. Jurnal Ilmiah Informatika Dan Ilmu Komputer (JIMA-ILKOM), 3(1), Article 1. https://doi.org/10.58602/jima-ilkom.v3i1.26

Joshi, M. A. P., & Patel, B. V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. Oriental Journal of Computer Science and Technology, 13(2,3), 78–81. https://doi.org/10.13005/ojcst13.0203.03

Leis, A. M., McSpadden, E., Segaloff, H. E., Lauring, A. S., Cheng, C., Petrie, J. G., Lamerato, L. E., Patel, M., Flannery, B., Ferdinands, J., Karvonen-Gutierrez, C. A., Monto, A., & Martin, E. T. (2023). K-medoids clustering of hospital admission characteristics to classify severity of influenza virus infection. Influenza and Other Respiratory Viruses, 17(3), e13120. https://doi.org/10.1111/irv.13120

Meiriza, A., Ali, E., Rahmiati, & Agustin. (2023). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Program BPJS Ketenagakerjaan. The Indonesian Journal of Computer Science, 12(2). https://doi.org/10.33022/ijcs.v12i2.3184

Momahhed, S. S., Emamgholipour Sefiddashti, S., Minaei, B., & Shahali, Z. (2023). K-means clustering of outpatient prescription claims for health insureds in Iran. BMC Public Health, 23(1), 788. https://doi.org/10.1186/s12889-023-15753-1

Ningrum, H., Irawan, E., & Lubis, M. R. (2021). Implementasi Metode K-Medoids Clustering Dalam Pengelompokan Data Penyakit Alergi Pada Anak. Jurasik (Jurnal Riset Sistem Informasi Dan Teknik Informatika), 6(1), Article 1. https://doi.org/10.30645/jurasik.v6i1.277

Nirwana, S. D., Jambak, M. I., & Bardadi, A. (2022). PERBANDINGAN ALGORITMA K-MEANS DAN K-MEDOIDS DALAM CLUSTERING RATA-RATA PENAMBAHAN KASUS COVID-19 BERDASARKAN KOTA/KABUPATEN DI PROVINSI SUMATERA SELATAN. JSiI (Jurnal Sistem Informasi), 9(2), Article 2. https://doi.org/10.30656/jsii.v9i2.5127

Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024). Analisis Supervised dan Unsupervised Learning pada Machine Learning: Systematic Literature Review. Jurnal Ilmu Komputer dan Informatika, 4(1), Article 1. https://doi.org/10.54082/jiki.168

Permana, I., & Salisah, F. N. S. (2022). Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation: The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm. Indonesian Journal of Informatic Research and Software Engineering (IJIRSE), 2(1), Article 1. https://doi.org/10.57152/ijirse.v2i1.311

Permatasari, R., Sukirman, & Fazza, F. E. (2024). Perbandingan Optimasi Penyuluhan Penyakit Stunting Pada Balita Integrasi Algoritma K- Means Dan Partitioning Around Medoids (PAM). Jurnal Ilmu Komputer Dan Teknologi Informasi, 1(2), Article 2. https://doi.org/10.71466/jiktif.v1i2.43

Purba, W. N., Sembiring, G. A., Turnip, M. T., Saputra, A., & Manihuruk, B. J. I. (2023). PENERAPAN DATA MINING UNTUK PENGELOLAAN DATA REKAM MEDIS MENGGUNAKAN METODE K-

MEANS CLUSTERING PADA RUMAH SAKIT ROYAL PRIMA MEDAN. Jurnal Teknik Informasi Dan Komputer (Tekinkom), 6(1), Article 1. https://doi.org/10.37600/tekinkom.v6i1.857

Razaki, A., Chrisnanto, Y. H., & Melina, M. (2024). Penanganan Outlier Pada Metode Algoritma K- Nearest Neighbors (KNN) Dengan Metode Kernel Density Estimation Pada Kasus Penyakit Diabetes. INTECOMS: Journal of Information Technology and Computer Science, 7(4), 1177–1188. https://doi.org/10.31539/intecoms.v7i4.10866

Romli, I. (2021). PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS UNTUK KLASIFIKASI PENYAKIT ISPA. Indonesian Journal of Business Intelligence (IJUBI), 4(1), Article 1. https://doi.org/10.21927/ijubi.v4i1.1727

Safitri, E. M. (2024). Clustering Study Of Hospitals In Bojonegoro Based On Health Workers With K-Means And K-Medoids Methods. Jurnal Statistika Dan Komputasi, 3(2), Article 2. https://doi.org/10.32665/statkom.v3i2.3592

Samosir, F. V. P., Mustamu, L. P., Anggara, E. D., Wiyogo, A. I., & Widjaja, A. (2021). Exploratory Data Analysis terhadap Kepadatan Penumpang Kereta Rel Listrik. Jurnal Teknik Informatika Dan Sistem Informasi, 7(2), Article 2. https://doi.org/10.28932/jutisi.v7i2.3700

Sari, Z. D. R., Arvita, Y., & Jasmir, J. (2024). Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5 Zudyanti Dwi Rahma Sari1, Ja. Jurnal Informatika Dan Rekayasa Komputer(JAKAKOM), 4(1), 827–834. https://doi.org/10.33998/jakakom.2024.4.1.1624

Setiawan, M. A. M., Kusrini, K., & Hartono, A. D. (2025). Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass. Jurnal Informatika: Jurnal Pengembangan IT, 10(1), Article 1. https://doi.org/10.30591/jpit.v10i1.8334

Syamfithriani, T. S., Mirantika, N., & Trisudarmo, R. (2023). Perbandingan Algoritma K-Means dan K-Medoids Untuk Pemetaan Daerah Penanganan Diare Pada Balita di Kabupaten Kuningan. Jurnal Sistem Informasi Bisnis, 12(2), 132–139. https://doi.org/10.21456/vol12iss2pp132-139

Utomo, W. (2021). The comparison of k-means and k-medoids algorithms for clustering the spread of the covid-19 outbreak in Indonesia. ILKOM Jurnal Ilmiah, 13(1), Article 1. https://doi.org/10.33096/ilkom.v13i1.763.31-35

Wahyudi, E. E., Auzan, M., Dharmawan, A., Nuryanto, D. E., Susyanto, N., Samodra, G., & Hadmoko, D. S. (2022). Akuisisi Data Prediksi Curah Hujan Secara Periodik Menggunakan Apache Airflow. Journal of Informatics Information System Software Engineering and Applications (INISTA), 4(2), Article 2. https://doi.org/10.20895/inista.v4i2.574