MIND (Multimedia Artificial Intelligent Networking Database) Journal

MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 10 | No. 1 | Halaman 99 - 113
DOI : https://doi.org/10.26760/mindjournal.v10i1.99-113 Juni 2025

Pendekatan *Unsupervised learning* dalam Segmentasi Kesehatan: Perbandingan K-Means dan DBSCAN

ANIS FITRI NUR MASRURIYAH, MARDIAH, MUHAMMAD DWI ANANDA, KARENINA NURMELITA MALIK

Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jakarta
Email: masruriyah@upnvj.ac.id

Received 17 Maret 2025 **|** *Revised* 23 Mei 2025 **|** *Accepted* 31 Mei 2025

ABSTRAK

Segmentasi kesehatan berbasis data pemeriksaan medis penting untuk mendukung strategi pencegahan penyakit. Penelitian ini membandingkan metode clustering K-Means dan DBSCAN menggunakan Silhouette Score dan Davies-Bouldin Index. Hasil menunjukkan bahwa K-Means dengan 8 cluster memberikan performa terbaik dengan Silhouette Score 0.2972 dan Davies-Bouldin Index 1.2934, dibandingkan konfigurasi lainnya. DBSCAN memperoleh Silhouette Score 0.2837, menunjukkan pendekatan berbasis densitas juga efektif dalam pengelompokan data. Dengan hasil ini, K-Means dengan 8 cluster dipilih sebagai metode terbaik untuk segmentasi kesehatan dalam penelitian ini. Temuan ini dapat mendukung analisis data medis untuk pencegahan penyakit yang lebih efektif dan personal.

***Kata kunci**: Segmentasi Kesehatan, Clustering, K-Means, DBSCAN, Silhouette Score, Davies-Bouldin Index*

ABSTRACT

Health segmentation based on medical examination data plays a crucial role in supporting disease prevention strategies. This study compares K-Means and DBSCAN clustering methods, evaluated using Silhouette Score and Davies-Bouldin Index, to identify the most effective segmentation approach. Experimental results indicate that K-Means with 8 clusters achieves the best performance, yielding a Silhouette Score of 0.2972 and a Davies-Bouldin Index of 1.2934, outperforming other configurations. Meanwhile, DBSCAN attains a Silhouette Score of 0.2837, demonstrating the efficacy of density-based clustering in handling medical data. Based on these findings, K-Means with 8 clusters emerges as the most optimal method for health segmentation in this study. These insights contribute to the advancement of data-driven disease prevention strategies and personalized healthcare management..

***Keywords**: Health Segmentation, Clustering, K-Means, DBSCAN, Silhouette Score, Davies-Bouldin Index*

1. PENDAHULUAN

Peningkatan kesadaran akan kesehatan masyarakat dan perkembangan teknologi di bidang kesehatan telah mendorong pemanfaatan data medis dalam pengambilan keputusan klinis dan kebijakan kesehatan **(Aram, 2021; Faisal et al., 2021; Yang et al., 2024)**. Pemeriksaan medis rutin menghasilkan data yang kaya akan informasi terkait kondisi fisiologis individu, seperti tekanan darah, kadar glukosa, indeks massa tubuh, dan kebiasaan hidup. Namun, pemanfaatan data ini masih cenderung berfokus pada diagnosis atau prediksi penyakit **(Masruriyah, Novita, & Sukmawati, 2023; Masruriyah, Novita, Sukmawati, Arif, et al., 2023; Masruriyah, Novita, Sukmawati, Fauzi, et al., 2023; Mia et al., 2022; Sonjaya et al., 2022),** strategi preventif atau upaya untuk melakukan pencegahan berdasarkan hasil pemeriksaan medis sering kali kurang mendapat perhatian. Oleh karena itu, penelitian ini bertujuan untuk menerapkan metode *clustering* guna melakukan segmentasi berbasis data pemeriksaan medis. Pemahaman pola segmentasi berbasis data pemeriksaan medis dapat bermanfaat untuk melakukan strategi preventif lebih cepat dan tepat sasaran sebelum penyakit berkembang lebih lanjut.

Berbagai penelitian telah mengusulkan pendekatan berbasis *machine learning* untuk pengelompokan dan prediksi risiko kesehatan. Antara lain pada penelitian Penafiel et al. (2021) yang mengembangkan model prediksi risiko stroke berbasis teori Dempster-Shafer yang mampu menangani data dengan missing values serta memberikan interpretasi yang lebih transparan bagi tenaga medis. Penelitian Choi et al. (2022) menggunakan *machine learning* untuk memprediksi pengeluaran medis tinggi, namun penelitian ini lebih berfokus pada aspek ekonomi daripada pengelompokan individu berdasarkan kondisi kesehatan. Sementara itu, Shpigelman dan Shamir (2023) mengembangkan algoritma FRIGATE untuk meningkatkan efektivitas *clustering* dalam data medis dengan metode seleksi fitur berbasis teori permainan. Namun, penelitian mereka lebih menitikberatkan pada optimasi algoritma dibandingkan implementasi dalam konteks segmentasi kesehatan.

Selain itu, penelitian oleh Johari et al. (2020) mengembangkan aplikasi berbasis decision tree untuk prediksi penyakit, tetapi masih belum mempertimbangkan segmentasi kesehatan sebagai langkah awal sebelum diagnosis dilakukan. Yang et al. (2024) menerapkan berbagai metode *clustering* dalam sistem smart healthcare, tetapi lebih difokuskan pada optimalisasi sumber daya medis ketimbang strategi pencegahan berbasis data. Sementara itu, Dubey et al. (2022) mengeksplorasi kombinasi algoritma *clustering* dengan optimasi berbasis Teaching-Learning-Based Optimization (TLBO), yang menunjukkan peningkatan akurasi dalam segmentasi data medis. Namun, penelitian ini masih kurang membahas interpretasi klinis dari hasil *clustering* yang diperoleh.

Meskipun penelitian-penelitian sebelumnya telah mengadopsi berbagai teknik *clustering* untuk analisis data medis, namun masih terdapat beberapa gap penelitian yang perlu diisi. Pertama, sebagian besar penelitian lebih menitikberatkan pada aspek prediktif daripada eksplorasi segmentasi kesehatan. Padahal, segmentasi kesehatan sangat penting untuk memahami kelompok individu dengan karakteristik kesehatan serupa guna merancang strategi pencegahan yang lebih efektif **(Nnoaham & Cann, 2020).** Kedua, banyak penelitian yang menggunakan *clustering* untuk optimasi sistem kesehatan atau perencanaan ekonomi, tetapi belum banyak yang secara eksplisit menekankan interpretasi klinis hasil *clustering* untuk membantu tenaga medis dalam pengambilan keputusan **(AL-Kahil et al., 2020; Aram, 2021; Meneses Navarro et al., 2022; Sugiura et al., 2024).**

Ketiga, sebagian besar studi hanya mengevaluasi performa algoritma *clustering* dari segi metrik kuantitatif, seperti akurasi dan indeks validasi *clustering*, tanpa menghubungkannya

dengan implikasi medis yang nyata. Evaluasi *clustering* seharusnya tidak hanya didasarkan pada performa matematis, tetapi juga harus mencerminkan kegunaannya dalam mendukung strategi kesehatan yang berbasis data. Oleh karena itu, penelitian ini berusaha untuk mengisi gap tersebut dengan menerapkan metode K-Means dan DBSCAN dalam segmentasi kesehatan berdasarkan data pemeriksaan medis.

Meskipun banyak penelitian sebelumnya menerapkan *clustering* pada data medis, masih jarang yang secara eksplisit membandingkan dua pendekatan berbeda seperti centroid-based (K-Means) dan density-based (DBSCAN) pada data pemeriksaan rutin yang mengandung kombinasi atribut kontinu dan kategorikal. Pemilihan kedua metode ini bertujuan untuk mengevaluasi keunggulan relatif masing-masing dalam mengidentifikasi kelompok risiko serta *outlier* klinis.

Dalam penelitian ini, algoritma K-Means dipilih karena kemampuannya dalam mengelompokkan data dengan jumlah *cluster* yang telah ditentukan secara eksplisit. Berdasarkan penelitian **(Costa et al., 2022)** Performa k-means lebih baik jika dibandingkan dengan k-prototype dan KAMILA dalam membagi cluster pada tipe data campuran, yaitu data yang tersusun atas data kategorikal dan numerik. Selain itu DBSCAN juga dipilih karena kemampuannya dalam mengidentifikasi pola berbasis densitas serta mendeteksi outlier yang sering ditemukan dalam data medis. Selanjutnya penelitian **(Sutramiani et al., 2024)** Algoritma DBSCAN dapat mengidentifikasi variasi data yang lebih kompleks dan secara efektif memisahkan titik-titik yang tidak termasuk dalam klaster utama (*outlier*). dengan membandingkan kedua algoritma ini, penelitian ini diharapkan dapat memberikan pemahaman lebih dalam tentang bagaimana metode *clustering* dapat digunakan untuk mengelompokkan individu berdasarkan kondisi kesehatan mereka, serta bagaimana hasil segmentasi ini dapat diinterpretasikan dalam konteks klinis.

Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam pengelolaan kesehatan berbasis data dengan mengidentifikasi kelompok individu yang memiliki risiko kesehatan serupa. Dengan adanya segmentasi yang jelas, tenaga medis dan pembuat kebijakan dapat lebih mudah merancang intervensi yang tepat sasaran untuk pencegahan penyakit dan peningkatan kualitas hidup. Oleh karena itu, penelitian ini tidak hanya berkontribusi dalam pengembangan metode *clustering* untuk data kesehatan, tetapi juga dalam penerapannya dalam strategi kesehatan preventif yang lebih efektif.

2. METODOLOGI

2.1. Data

Dataset yang digunakan pada penelitian ini merujuk pada **Hasan (2025)** yang mencakup informasi dari individu yang menjalani pemeriksaan medis, dengan fokus pada faktor fisiologis dan gaya hidup yang berpengaruh terhadap kesehatan. Data ini tersusun atas tipe data kategorikal dan numerik, terdiri dari berbagai atribut yang mencerminkan kondisi kesehatan pasien, seperti tekanan darah, kadar glukosa, kebiasaan merokok, konsumsi alkohol, serta aktivitas fisik. Detail atribut ditunjukkan pada Tabel 1.

Tabel 1 Deskripsi Dataset

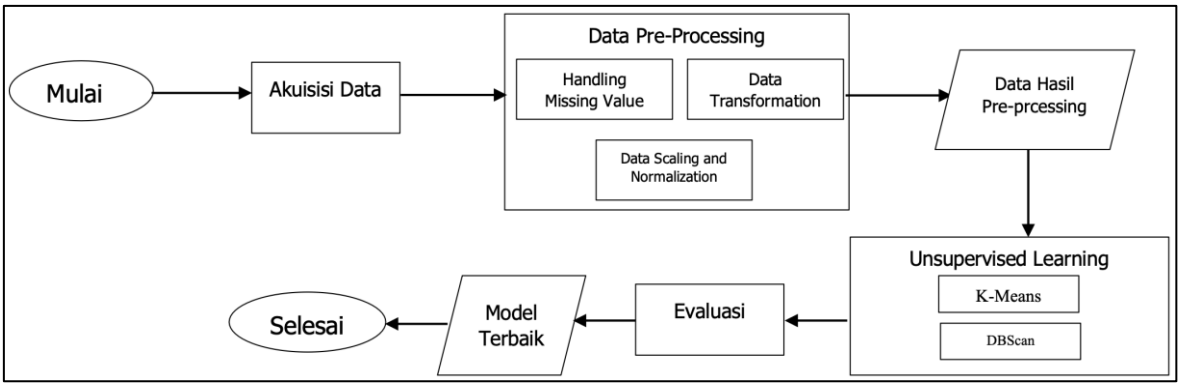
No.	Atribut	Keterangan
1	*Age*	Menunjukkan usia individu dalam hari, yang merupakan faktor utama dalam analisis risiko kesehatan.
2	*Sex*	Membedakan jenis kelamin individu, yang berperan dalam perbedaan fisiologis dan risiko penyakit.
3	*Height*	Mengukur tinggi badan dalam satuan sentimeter, yang digunakan untuk menghitung indeks massa tubuh (BMI) sebagai indikator kesehatan tubuh.
4	*Weight*	Berat badan individu dalam kilogram, yang bersama dengan tinggi badan digunakan untuk mengevaluasi status gizi dan risiko obesitas.
5	*Ap_Hi*	Tekanan darah sistolik yang mencerminkan kekuatan aliran darah terhadap dinding arteri saat jantung berkontraksi.
6	*Ap_Lo*	ekanan darah diastolik yang mengukur tekanan dalam arteri saat jantung beristirahat di antara kontraksi.
7	*Cholesterol*	Kadar kolesterol dalam darah
8	*Gluc*	Kadar glukosa darah yang menunjukkan risiko diabetes mellitus
9	*Smoke*	Status kebiasaan merokok individu
10	*Alco*	Konsumsi alkohol yang dapat mempengaruhi tekanan darah, fungsi hati, serta meningkatkan risiko penyakit kardiovaskular dan metabolik.
11	*Active*	Tingkat aktivitas fisik individu.
12	*Cardio*	Aktivitas fisik yang meningkatkan detak jantung dan pernapasan.

Dataset ini relevan untuk penelitian karena memiliki struktur data yang jelas, variabel yang mencerminkan faktor risiko kesehatan yang terukur, serta jumlah observasi yang cukup untuk menghasilkan analisis yang valid. Dengan adanya atribut-atribut kesehatan yang komprehensif, dataset ini memungkinkan eksplorasi mendalam dalam segmentasi kesehatan, yang dapat membantu dalam pengelompokan individu berdasarkan karakteristik medis dan kebiasaan hidup mereka.

2.2. Metode Penelitian

Penelitian ini menerapkan pendekatan *unsupervised learning* untuk melakukan segmentasi kesehatan berdasarkan data pemeriksaan medis. Langkah-langkah pada penelitian ini ditunjukkan pada Gambar 1, dimulai dari Akuisisi Data, dilanjutkan dengan Data Pre-Processing yang mencakup penanganan nilai hilang (*handling missing values*), transformasi data (*data transformation*), serta normalisasi dan penskalaan adata (data scaling). Setelah data dipraproses (data preprocessing), langkah berikutnya adalah penerapan metode *Unsupervised learning*, yaitu K-Means dan DBScan. Hasil dari proses *clustering* ini dievaluasi menggunakan Davies-Bouldin Index dan Silhouette score, kemudian model terbaik dipilih berdasarkan hasil evaluasi.

Proses analisis data dilakukan secara sistematis, dimulai dari tahap akuisisi data hingga evaluasi model guna menentukan metode *clustering* yang paling optimal. Akuisisi data merupakan langkah awal dalam penelitian ini, di mana dataset dikumpulkan dari sumber *open-source* yang kredibel.



Gambar 1. Alur Penelitian

Setelah data diperoleh, tahap data pre-processing dilakukan untuk memastikan kualitas, konsistensi, dan kesiapan data sebelum dianalisis lebih lanjut. Proses ini mencakup penanganan nilai yang hilang dengan metode yang tepat guna menjaga integritas data. Selain itu, dilakukan transformasi data untuk menyelaraskan format dan struktur dengan kebutuhan algoritma *clustering*. Transformasi ini mencakup konversi usia dari satuan hari ke tahun untuk meningkatkan interpretabilitas medis, pengkodean variabel kategorikal seperti "gluc" dan "cholesterol" menggunakan *one-hot encoding* guna menghindari asumsi ordinal yang tidak tepat, serta normalisasi atau standardisasi atribut numerik agar seluruh fitur berada pada skala yang seragam. Langkah-langkah ini penting untuk mencegah dominasi fitur tertentu dalam perhitungan jarak dan memastikan performa algoritma *clustering* yang optimal.

Karena data mencakup atribut campuran (numerik dan kategorikal biner), semua data dikonversi dan diskalakan menggunakan StandardScaler setelah proses encoding. Evaluasi mempertimbangkan keterbatasan Euclidean Distance dalam menangani atribut biner, namun metrik ini tetap digunakan untuk menjaga konsistensi dengan karakteristik algoritma dasar yang dibandingkan dalam penelitian ini.

Tahap berikutnya adalah penerapan algoritma *clustering* menggunakan K-Means dan DBScan. K-Means digunakan untuk mengelompokkan data berdasarkan kedekatan *centroid*, di mana jumlah *cluster* harus ditentukan sebelumnya **(Ahmed et al., 2020; Hairani et al., 2020; Ikotun et al., 2023; Setiawati et al., 2024)**. Proses dimulai dengan menentukan jumlah *cluster* optimal menggunakan Elbow Method, di mana nilai k dipilih berdasarkan titik siku pada grafik Within-*Cluster* Sum of Squares (WCSS). Setelah k ditentukan, algoritma secara iteratif menginisialisasi *centroid* secara acak, mengelompokkan data berdasarkan jarak terdekat ke *centroid*, dan memperbarui posisi *centroid* hingga konvergen. Secara sederhana cara kerja K-Means ditunjukkan pada Pseudocode 1.

Pseudocode 1 K-Means

Initialize k *centroid*s randomly
Repeat until convergence:
  Assign each data point to the nearest *centroid*
  Update *centroid*s by computing the mean of assigned points
  If *centroid*s do not change significantly, stop

Sebaliknya, DBScan mengelompokkan data berdasarkan kepadatan titik dalam ruang data, dengan mendefinisikan eps (radius maksimal untuk menganggap titik sebagai tetangga) dan minPts (jumlah minimum titik dalam radius tersebut untuk membentuk *cluster*) **(de la Selle et al., 2024; Gunawan, 2021; Setiawati et al., 2024)**. DBScan mampu mendeteksi outlier

sebagai *noise*, sehingga lebih fleksibel dibandingkan K-Means dalam menangani data dengan bentuk distribusi yang kompleks. Cara kerja DBScan ditunjukkan pada Pseudocode 2.

Pseudocode 2 DBScan

```
For each point P in dataset:
    If P is unvisited:
        Mark P as visited
        Retrieve neighbors within radius eps
        If number of neighbors ≥ minPts:
            Form a new cluster
            Expand cluster by recursively adding density-reachable points
        Else:
            Mark P as noise
```

Setelah kedua metode diterapkan, dilakukan proses evaluasi menggunakan metrik Silhouette Score dan Davies-Bouldin Index. Metrik ini digunakan untuk mengukur kualitas *clustering* dengan menilai seberapa baik data dalam satu *cluster* dikelompokkan serta seberapa jauh antar *cluster* yang terbentuk. Silhouette Score mengevaluasi seberapa baik suatu data dikelompokkan dalam sebuah *cluster* dengan membandingkan jarak rata-rata antara suatu titik dengan titik-titik dalam *cluster* yang sama (cohesion) dan jarak rata-rata dengan titik-titik dalam *cluster* terdekat lainnya (separation) **(Masruriyah et al., 2024).** Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa titik data berada dalam *cluster* yang tepat. Evaluasi ini digunakan untuk menentukan jumlah *cluster* yang optimal dengan memilih konfigurasi *clustering* yang menghasilkan Silhouette Score tertinggi. Pseudocode 3 menunjukkan cara kerja Silhoutte Score.

Pseudocode 3 Silhoutte Score

```
For each data point i:
    Compute a(i) = average distance of i to all points in the same cluster
    Compute b(i) = lowest average distance of i to all points in another cluster
    Compute silhouette(i) = (b(i) - a(i)) / max(a(i), b(i))
Compute overall Silhouette Score = mean(silhouette(i) for all i)
```

Selanjutnya, Davies-Bouldin Index (DBI) menilai kualitas *clustering* berdasarkan rasio antara dispersi dalam-*cluster* dan jarak antar-*cluster* **(Masruriyah et al., 2024).** Untuk setiap *cluster*, indeks ini mengukur seberapa mirip *cluster* tersebut dengan *cluster* lainnya menggunakan rasio antara variabilitas dalam-*cluster* dengan jarak antar-*cluster*. Nilai DBI yang lebih rendah menunjukkan bahwa *cluster* yang dihasilkan lebih terpisah dengan baik dan memiliki kompaksi yang lebih tinggi, sehingga *clustering* dianggap lebih optimal. Cara kerja DBI ditunjukkan pada Pseudocode 4.

Pseudocode 4 DBI

```
For each cluster i:
    Compute intra-cluster scatter S(i)
    For each other cluster j:
        Compute inter-cluster distance D(i, j)
    Compute R(i) = max((S(i) + S(j)) / D(i, j))
Compute DBI = mean(R(i) for all clusters i)
```

Pada tahap akhir, model dengan performa terbaik dipilih berdasarkan hasil evaluasi. Model yang terpilih diharapkan dapat memberikan segmentasi kesehatan yang representatif serta membantu dalam memahami pola kesehatan populasi. Sehingga, penelitian ini berkontribusi

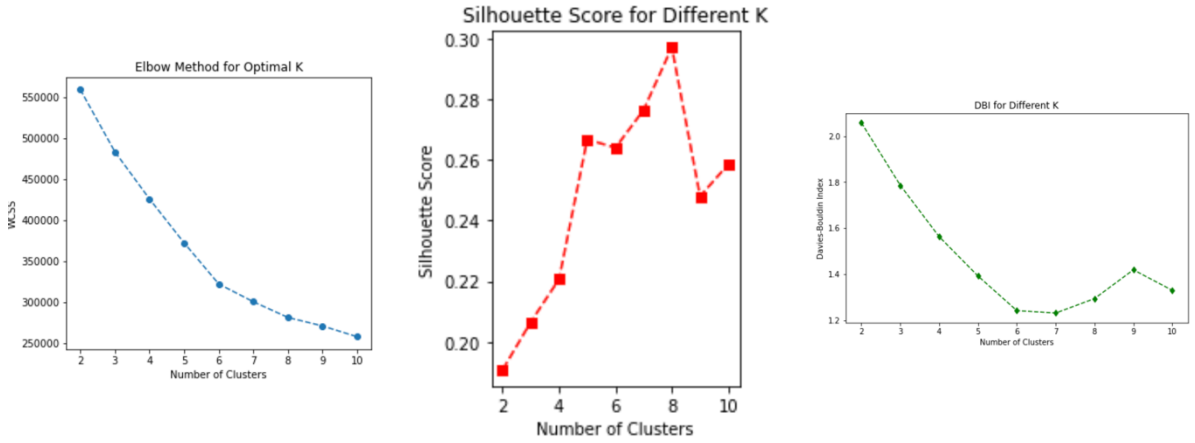
dalam menyediakan wawasan berbasis data yang dapat mendukung strategi pencegahan penyakit secara lebih efektif.

3. HASIL DAN PEMBAHASAN

Dalam dataset yang digunakan, atribut usia (*age*) awalnya direpresentasikan dalam satuan hari. Meskipun informasi dalam bentuk hari lebih granular, penggunaan usia dalam tahun lebih intuitif dan relevan untuk analisis medis. Sebagian besar penelitian kesehatan dan praktik medis menggunakan usia dalam tahun karena lebih mudah diinterpretasikan dalam kaitannya dengan risiko kesehatan dan pola penyakit. Selain itu, perbedaan usia dalam hitungan hari tidak memberikan nilai tambah yang signifikan dalam analisis *clustering*, sedangkan perbedaan usia dalam tahun dapat membantu dalam memahami distribusi kelompok umur yang lebih jelas. Oleh karena itu, konversi dari hari ke tahun dilakukan dengan membagi nilai usia dengan 365, sehingga menghasilkan representasi yang lebih komprehensif dan informatif. Misal pada baris pertama data tertera usia 18393 hari, sehingga transformasi menjadi 50 tahun.

Selanjutnya, beberapa atribut dalam dataset, seperti gluc (1,2,3) dan cholesterol (1,2,3), memiliki nilai kategori yang menunjukkan tingkat dari rendah ke tinggi. Meskipun angka ini tampak seperti skala numerik, mereka sebenarnya adalah representasi kategorikal yang tidak memiliki makna ordinal yang jelas dalam konteks perhitungan jarak dalam algoritma *clustering* seperti K-Means dan DBSCAN. Jika dibiarkan dalam bentuk angka asli, algoritma akan menganggapnya sebagai variabel kontinu, yang dapat mengarah pada kesalahan dalam perhitungan jarak Euclidean. Oleh karena itu, variabel-variabel ini dikonversi menggunakan one-hot encoding, di mana setiap kategori direpresentasikan sebagai variabel biner terpisah, misalnya gluc_1, gluc_2, gluc_3.

Selanjutnya, berdasarkan Gambar 2 hasil eksperimen menunjukkan bahwa pemilihan jumlah klaster dalam metode K-Means berdasarkan metode Elbow (a), Silhouette Score (b), dan Davies-Bouldin Index (DBI) (c) memberikan indikasi yang tidak sepenuhnya konsisten. Metode Elbow menghasilkan grafik yang samar dengan indikasi titik siku yang muncul pada klaster ke-6, sementara grafik Silhouette Score menunjukkan pola fluktuatif, dan Davies-Bouldin Index menunjukkan adanya lembah dan puncak pada jumlah klaster 6 dan 9. Berdasarkan analisis ini, pemodelan K-Means dilakukan dengan jumlah klaster 5, 6, 7, 8, dan 9 untuk mengeksplorasi distribusi optimal dari data yang digunakan.



(a) **(b)** **(c)**

Gambar 2. Penentuan Jumlah Klaster

Dari hasil evaluasi, K-Means dengan 8 klaster memberikan nilai Silhouette Score tertinggi, yaitu 0.2972, yang menunjukkan bahwa data dalam klaster relatif lebih terpisah dari klaster lain dan memiliki kohesi yang lebih baik dibandingkan jumlah klaster lainnya. Namun, nilai Davies-Bouldin Index untuk K-Means 8 adalah 1.2934, yang meskipun lebih rendah dari beberapa klaster lainnya, masih berada dalam kisaran yang menunjukkan adanya tumpang tindih antar klaster. Sebagai perbandingan, K-Means dengan 7 klaster memiliki nilai Davies-Bouldin Index paling rendah, yaitu 1.2305, yang mengindikasikan pemisahan klaster yang relatif lebih baik, meskipun nilai Silhouette Score (0.2765) sedikit lebih rendah dibandingkan K-Means dengan 8 klaster. Sementara itu, K-Means dengan 9 klaster memiliki Davies-Bouldin Index tertinggi (1.4185) dan Silhouette Score terendah (0.2480), yang menunjukkan bahwa penggunaan 9 klaster menghasilkan pemisahan klaster yang kurang optimal.

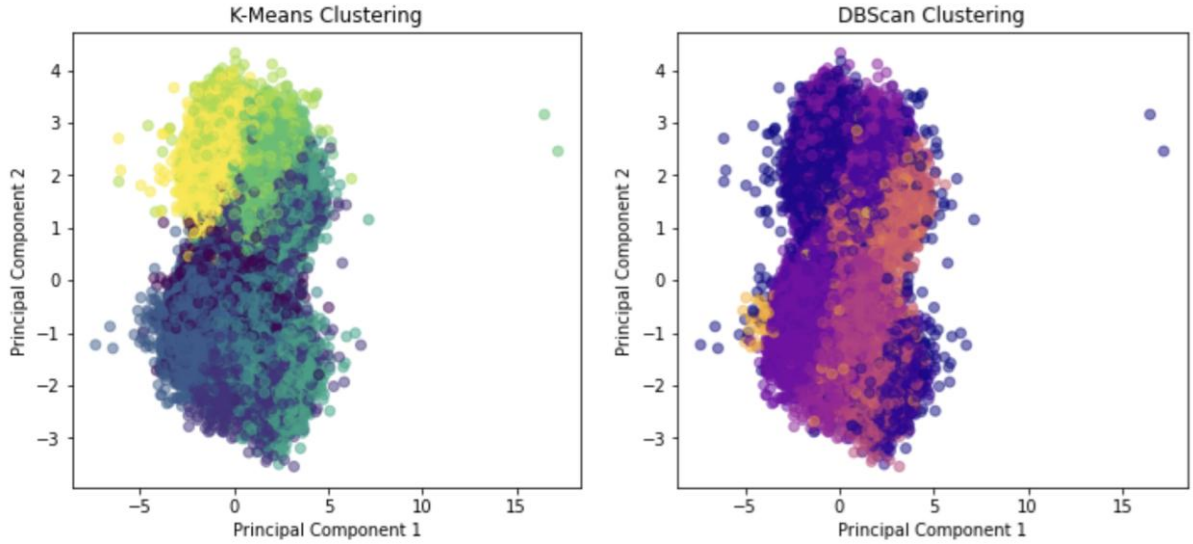
Pada metode DBScan, diperoleh Silhouette Score sebesar 0.2837 dengan jumlah outlier yang terdeteksi sebanyak 506 data. Nilai ini lebih tinggi dibandingkan beberapa model K-Means, menunjukkan bahwa DBScan mampu membentuk klaster yang cukup baik berdasarkan densitas, dengan identifikasi outlier yang cukup signifikan. Deteksi outlier ini penting karena menunjukkan adanya sampel data yang tidak sesuai dengan pola umum dalam klaster utama, yang mungkin tidak dapat diidentifikasi dengan metode berbasis *centroid* seperti K-Means.

Dari hasil ini, dapat disimpulkan bahwa metode pengelompokan yang paling baik bergantung pada perspektif evaluasi yang digunakan. Jika mempertimbangkan keseimbangan antara kohesi klaster dan pemisahan antar-klaster, K-Means dengan 7 atau 8 klaster dapat menjadi pilihan yang lebih baik. Namun, jika tujuan utama adalah mendeteksi outlier dan menemukan struktur klaster berdasarkan kepadatan data, DBScan memberikan keunggulan karena mampu mengidentifikasi data yang tidak sesuai dengan pola mayoritas. Oleh karena itu, dalam konteks analisis data medis ini, kombinasi kedua pendekatan dapat memberikan wawasan yang lebih komprehensif tentang pola tersembunyi dalam dataset

Dengan pendekatan ini, setiap kategori menjadi atribut independen yang tidak memiliki hubungan numerik langsung dengan kategori lainnya. Hal ini memastikan bahwa algoritma *clustering* dapat bekerja secara optimal tanpa bias yang disebabkan oleh asumsi hubungan matematis antar kategori. Transformasi ini juga membantu dalam interpretasi hasil *clustering*, karena dapat menunjukkan kecenderungan kelompok berdasarkan tingkat glukosa atau kolesterol tertentu.

Berdasarkan visualisasi hasil *clustering* pada Gambar 3, terdapat beberapa perbedaan mendasar antara metode K-Means (a) dan DBScan (b), yang dapat dianalisis dari pola distribusi klaster serta performa metrik evaluasi yang telah dihitung sebelumnya.

Pada hasil K-Means dengan K=8 (a), klaster yang terbentuk tampak lebih tersegmentasi dengan warna yang relatif terpisah, meskipun ada beberapa tumpang tindih antar titik. Hal ini mencerminkan bahwa K-Means mampu membagi data ke dalam kelompok tertentu berdasarkan kedekatan euclidean, yang sesuai dengan sifat algoritma ini yang mengasumsikan bahwa klaster berbentuk sferis. Dari hasil evaluasi, Silhouette Score untuk K=8 adalah 0.2972, yang merupakan skor tertinggi dibandingkan jumlah klaster lainnya. Ini menunjukkan bahwa data dalam klaster cukup kohesif dan memiliki pemisahan yang relatif baik. Namun, Davies-Bouldin Index (DBI) (b) sebesar 1.2934, yang menunjukkan bahwa meskipun pemisahan antar klaster cukup baik, ada beberapa klaster yang masih memiliki kedekatan tinggi satu sama lain.



K-Means (a) **DBScan (b)**

Gambar 3. Visualisasi Klaster

Sementara itu, hasil *clustering* dengan DBScan (b) menunjukkan pola klaster yang lebih bebas, tanpa harus membentuk grup berbentuk bola seperti pada K-Means. DBScan lebih cocok untuk data dengan kepadatan yang bervariasi karena mampu mendeteksi outlier secara langsung. Hal ini terlihat dari adanya titik-titik dengan warna berbeda yang tersebar di luar area utama, yang merupakan 506 *outlier* yang terdeteksi. Silhouette Score DBScan adalah 0.2837, yang cukup kompetitif dibandingkan dengan K-Means, menandakan bahwa metode ini berhasil mengelompokkan data dengan cukup baik berdasarkan kepadatan. Keunggulan utama DBScan dalam penelitian ini adalah kemampuannya mengidentifikasi *noise* atau data yang tidak sesuai dengan pola utama, yang tidak dapat dilakukan oleh K-Means.

Agar dapat memberikan pemahaman yang lebih mendalam terhadap hasil segmentasi, dilakukan analisis centroid dari masing-masing klaster K-Means dengan k = 8. Rata-rata nilai tiap fitur dalam setiap klaster disajikan pada Tabel 2.

Tabel 2 Rata-rata Setiap Atribut per Klaster (*Centroid*)

cluster	height	weight	ap_hi	ap_lo	age_years	male	female
0	169,92	76,5	129,39	100,13	52,61	0	1
1	161,3	69,72	118,25	80,39	51,01	1	0
2	160,98	74,44	139,17	109,71	54,46	1	0
3	167,67	78,27	129,72	105,13	51,94	0,31	0,69
4	163,23	78,3	134,04	103,22	53,53	0,73	0,27
5	163,58	77,28	133,49	100,2	55,16	0,72	0,28
6	169,27	77,06	127,8	97,16	51,81	0,14	0,86
7	164,16	73,61	127,31	95,11	52,74	0,69	0,31

cluster	cholesterol_normal	cholesterol_above_normal	cholesterol_well_above_normal	gluc_normal	gluc_above_normal	gluc_well_above_normal	smoke_yes
0	0,84	0,1	0,07	1	0	0	0
1	0,87	0,1	0,03	1	0	0	0
2	0,72	0,15	0,13	1	0	0	0
3	0,67	0,19	0,13	0,82	0,11	0,07	0,49
4	0,43	0,48	0,1	0	1	0	0,01
5	0,27	0,07	0,67	0	0	1	0,03
6	0,76	0,16	0,08	0,88	0,08	0,04	1
7	0,83	0,11	0,05	1	0	0	0

cluster	smoke_no	alco_yes	alco_no	active_yes	active_no	cardio_yes	cardio_no
0	1	0	1	1	0	0,5	0,5
1	1	0	1	1	0	0	1
2	1	0	1	1	0	1	0
3	0,51	1	0	0,85	0,15	0,48	0,52
4	0,99	0	1	0,78	0,22	0,59	0,41
5	0,97	0	1	0,8	0,2	0,63	0,37
6	0	0	1	0,83	0,17	0,48	0,52
7	1	0	1	0	1	0,52	0,48

Berdasarkan Tabel 2 dari setiap klaster K-Means dengan k = 8, diperoleh karakteristik yang cukup berbeda dan dapat diinterpretasikan dalam konteks medis. Interpretasi tiap klaster dimulai dari Klaster 1 yang menunjukkan profil paling sehat di antara seluruh klaster. Tekanan darah sistolik dan diastolik berada pada nilai rendah (118.25 dan 80.39), usia rata-rata 51.01 tahun, dan seluruh individu dalam klaster ini aktif secara fisik, tidak merokok maupun mengonsumsi alkohol. Tidak ada kasus penyakit kardiovaskular tercatat pada klaster ini (cardio_yes = 0). Klaster ini dapat dianggap sebagai representasi kelompok sehat dan ideal dalam konteks preventif. Selanjutnya, klaster 2 memperlihatkan profil dengan risiko tertinggi. Usia rata-rata tertua (54.46 tahun), tekanan darah tertinggi (139.17/109.71), dan seluruh individu mengalami kondisi kardiovaskular (cardio_yes = 1). Selain itu, kadar glukosa dan kolesterol juga tergolong tinggi. Ini menunjukkan bahwa klaster ini merupakan kelompok yang secara klinis berada dalam risiko tinggi, terutama untuk hipertensi dan penyakit metabolik.

Pada Klaster 3 dicirikan oleh pola gaya hidup yang buruk. Sekitar 49% anggotanya adalah perokok dan seluruhnya mengonsumsi alkohol. Meskipun nilai tekanan darah dan kolesterol tidak setinggi klaster 2, pola perilaku ini menjadikan kelompok ini penting untuk pendekatan intervensi berbasis perubahan gaya hidup. Aktivitas fisik relatif cukup (active_yes = 0.85), namun tidak mengimbangi faktor risikonya. Kemudian, Klaster 4 memiliki kombinasi tekanan darah tinggi (134.04/103.22), usia relatif lanjut (53.53), dan kadar kolesterol yang tidak ideal. Sekitar 59% mengalami masalah kardiovaskular. Klaster ini bisa dikategorikan sebagai kelompok risiko menengah menuju tinggi, dan relevan untuk intervensi berbasis pemeriksaan berkala.

Klaster 5 sangat menonjol dari sisi kadar glukosa. Seluruh individu dalam klaster ini memiliki kadar glukosa sangat tinggi (*gluc_well_above_normal* = 1.00), yang menunjukkan kemungkinan besar mengidap diabetes atau berada pada kondisi pra-diabetes berat. Kondisi kardiovaskular juga cukup dominan (63%). Kelompok ini sangat penting untuk pendekatan preventif yang fokus pada penyakit metabolik. Di sisi lain, Klaster 6 didominasi oleh perempuan (86%), memiliki tekanan darah yang lebih rendah, dan tingkat aktivitas fisik yang tinggi (83%). Risiko kardiovaskular relatif sedang (cardio_yes = 48%), menjadikan klaster ini sebagai populasi aktif namun tetap perlu pemantauan.

Klaster 7 memperlihatkan nilai tekanan darah yang moderat (127.31/95.11), dan proporsi laki-laki cukup besar (69%). Kondisi kesehatan secara umum stabil, dengan risiko kardiovaskular yang hampir seimbang (52%). Terakhir, Klaster 0 menunjukkan profil kesehatan yang moderat. Tekanan darah sedikit tinggi, usia rata-rata 52.61, dan separuh populasi menunjukkan gejala kardiovaskular. Kelompok ini mungkin termasuk populasi yang berada di ambang risiko dan dapat memperoleh manfaat besar dari intervensi dini. Dengan memahami karakteristik masing-masing klaster ini, hasil *clustering* tidak hanya dapat digunakan untuk pengelompokan teknis, tetapi juga dapat dihubungkan dengan program kesehatan publik seperti deteksi dini, skrining rutin, dan edukasi gaya hidup sehat yang lebih tepat sasaran.

Berdasarkan hasil interpretasi centroid, K-Means dengan K=8 tidak hanya memberikan hasil terbaik secara kuantitatif melalui nilai Silhouette Score tertinggi, tetapi juga menghasilkan segmentasi yang bermakna secara klinis. Masing-masing klaster menunjukkan pola kesehatan yang berbeda: mulai dari kelompok sehat tanpa risiko (Klaster 1), kelompok dengan risiko tinggi hipertensi dan penyakit kardiovaskular (Klaster 2), hingga kelompok dengan pola hidup tidak sehat (Klaster 3) dan risiko metabolik tinggi (Klaster 5).

Sementara itu, DBSCAN tetap relevan karena berhasil mengidentifikasi 506 outlier, yang dapat mencerminkan pasien dengan kondisi medis unik atau komorbiditas yang tidak sesuai dengan

pola mayoritas. Kemampuan ini menjadikan DBSCAN relevan dalam mengidentifikasi anomali medis yang berpotensi signifikan dalam praktik klinis.

Hasil klasterisasi menggunakan algoritma DBSCAN menghasilkan 29 klaster (Tabel 3), yang diberi label mulai dari klaster 0 hingga 28, seperti yang ditunjukkan pada Tabel 3. Setiap klaster menunjukkan karakteristik yang unik berdasarkan atribut usia, jenis kelamin, tekanan darah, kadar kolesterol, kadar glukosa, serta gaya hidup. Klaster 0 terdiri dari individu berusia sekitar 50 tahun, baik laki-laki maupun perempuan, dengan kadar kolesterol normal dan glukosa bervariasi antara normal dan tidak normal. Gaya hidup yang dijalani umumnya sehat, ditandai dengan tidak merokok, tidak mengonsumsi alkohol, dan aktif secara fisik, sehingga kelompok ini dikategorikan relatif sehat meskipun terdapat indikasi gangguan glukosa.

Klaster 1 menunjukkan karakteristik berisiko tinggi, terdiri dari laki-laki dengan usia rata-rata 56 tahun, tekanan darah dan kadar kolesterol tinggi, meskipun kadar glukosa tergolong normal. Para anggota klaster ini tetap menjalani gaya hidup aktif. Klaster 2 dan 3 mengelompokkan individu berdasarkan jenis kelamin, di mana klaster 2 berisi laki-laki dan klaster 3 berisi perempuan. Kedua klaster ini memiliki tekanan darah tinggi (ap_hi > 133,9 dan ap_lo > 84) dengan kadar glukosa normal dan gaya hidup yang cukup sehat, yaitu tidak merokok dan aktif. Nilai tekanan darah yang tinggi mengindikasikan risiko pre-hipertensi.

Tabel 3 Ringkasan Klaster Pada DBSCAN

cluster	height	weight	ap_hi	ap_lo	age_years	male	female
0	169,79	73,52	120,28	78,62	50,96	0	1
1	160,01	76,89	135,34	85,11	56,12	1	0
2	159,92	78,5	133,95	84,72	56,57	1	0
3	169,5	77,17	132,96	84,29	54,03	0	1
4	162,14	68,97	117,93	77,14	50,79	1	0
5	159,4	75,44	130,16	83,36	55,64	1	0
6	161,61	69,21	117,54	76,88	50,76	1	0
7	169,69	73,52	119,93	79,05	50,57	0	1
.							
.							
.							
28	179	99	116	78	48	0	1

cluster	cholesterol_normal	cholesterol_above_normal	cholesterol_well_above_normal	gluc_normal	gluc_above_normal	gluc_well_above_normal	smoke_yes
0	1	0	0	1	0	0	0
1	0	0	1	1	0	0	0
2	0	0	1	1	0	0	0
3	1	0	0	1	0	0	0
4	1	0	0	1	0	0	0
5	0	0	1	1	0	0	0
6	1	0	0	1	0	0	0
7	1	0	0	1	0	0	0
.							
.							
.							
28	0	0	1	1	0	0	0

cluster	smoke_no	alco_yes	alco_no	active_yes	active_no	cardio_yes	cardio_no
0	1	0	1	1	0	0	1
1	1	0	1	1	0	1	0
2	1	0	1	0	1	1	0
3	1	0	1	1	0	1	0
4	1	0	1	0	1	0	1
5	1	0	1	1	0	0	1
6	1	0	1	1	0	0	1
7	1	0	1	0	1	0	1
.							
.							
.							
28	1	0	1	0	1	0	1

Selanjutnya, klaster 4 dan 5 terdiri dari laki-laki dengan kadar kolesterol dan glukosa yang relatif normal, namun dengan tekanan darah yang berada pada kisaran pre-hipertensi (klaster 4: ap_hi = 117,93 dan ap_lo = 77,14; klaster 5: ap_hi = 130,16 dan ap_lo = 83,36). Perbedaan utama antara kedua klaster ini terletak pada kadar kolesterol, di mana klaster 4 memiliki kadar kolesterol normal sementara klaster 5 berada di atas normal, sehingga individu pada klaster 5 disarankan untuk lebih memperhatikan asupan kolesterol.

Klaster 6 dan 7 mencakup laki-laki dan perempuan berusia sekitar 50 tahun, dengan kadar glukosa dan kolesterol normal serta tidak merokok. Perbedaan antara kedua klaster ini terdapat pada gaya hidup, di mana anggota klaster 6 cenderung aktif, sedangkan klaster 7 tidak aktif. Sementara itu, klaster 28 terdiri dari perempuan dengan usia termuda, yaitu 48 tahun, memiliki kadar kolesterol tinggi dan glukosa normal, serta menjalani gaya hidup tidak aktif. Meskipun masih tergolong muda, individu dalam klaster ini perlu meningkatkan aktivitas fisik dan memperhatikan kadar kolesterol mereka.

Secara keseluruhan, klaster 0, 6, dan 7 dapat dikategorikan sebagai kelompok yang paling sehat, ditandai dengan usia sekitar 50 tahun, tekanan darah, kolesterol, dan glukosa yang berada dalam rentang normal, meskipun terdapat variasi dalam gaya hidup. Di sisi lain, klaster 1 menunjukkan risiko tertinggi akibat usia yang lebih lanjut dan tekanan darah serta kolesterol yang tinggi. Klaster 2 hingga 5 dikategorikan sebagai kelompok dengan risiko menengah, sementara klaster 28 perlu mendapatkan perhatian khusus terhadap gaya hidup dan pengendalian kadar kolesterol. Temuan ini menunjukkan bagaimana algoritma DBSCAN mampu mengidentifikasi kelompok-kelompok populasi dengan karakteristik kesehatan yang berbeda secara signifikan berdasarkan data hasil pemeriksaan medis.

Dengan demikian, kedua pendekatan ini tidak bersifat saling eksklusif, melainkan saling melengkapi. K-Means memberikan segmentasi populasi yang jelas dan interpretable, sedangkan DBSCAN menambahkan nilai melalui deteksi kasus ekstrem. Kombinasi keduanya memberikan gambaran menyeluruh terhadap struktur data medical check-up, yang dapat dimanfaatkan untuk mendukung strategi kesehatan preventif yang lebih presisi dan personal.

4. KESIMPULAN

Penelitian ini membuktikan bahwa algoritma K-Means dengan jumlah klaster 8 mampu menghasilkan segmentasi kesehatan yang tidak hanya optimal secara metrik evaluasi (Silhouette Score tertinggi), tetapi juga bermakna secara klinis. Analisis centroid menunjukkan perbedaan karakteristik yang jelas di setiap klaster, mulai dari kelompok individu sehat tanpa risiko, hingga kelompok dengan tekanan darah tinggi, kadar glukosa ekstrem, dan gaya hidup tidak sehat. Segmentasi ini dapat menjadi dasar dalam merancang strategi pencegahan penyakit yang lebih tepat sasaran dan personal.

Sementara itu, algoritma DBSCAN memberikan kontribusi penting dalam mengidentifikasi 506 outlier yang tidak terdeteksi oleh K-Means. Temuan ini menunjukkan bahwa pendekatan berbasis densitas memiliki keunggulan dalam mendeteksi pola data tidak lazim, yang mungkin mewakili kasus medis unik atau kompleks. Oleh karena itu, penggunaan kedua metode secara bersamaan memberikan pemahaman yang lebih komprehensif terhadap struktur data pemeriksaan medis dan dapat mendukung keputusan kesehatan berbasis data secara lebih presisi.

Pada penelitian selanjutnya disarankan untuk mengeksplorasi algoritma *clustering* tambahan serta menggunakan metrik jarak yang lebih adaptif terhadap data campuran. Evaluasi berbasis validasi klinis nyata, integrasi visualisasi non-linear seperti t-SNE atau UMAP, serta studi longitudinal terhadap klaster berisiko tinggi dapat meningkatkan relevansi dan dampak praktis dari segmentasi ini dalam sistem kesehatan preventif.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Pembangunan Nasional "Veteran" Jakarta atas dukungan dan fasilitas yang telah diberikan sehingga penelitian ini dapat berjalan dengan baik. Dukungan institusi ini, baik dalam bentuk sumber daya, bimbingan akademik, maupun kesempatan untuk mengembangkan penelitian di bidang unsupervised learning pada data medis, sangat berperan dalam keberhasilan studi ini. Semoga hasil penelitian ini dapat memberikan kontribusi bagi perkembangan ilmu pengetahuan serta membuka peluang untuk penelitian lebih lanjut di masa depan.

DAFTAR RUJUKAN

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Issue 8). https://doi.org/10.3390/electronics9081295

AL-Kahil, A. B., Khawaja, R. A., Kadri, A. Y., Abbarh,MBBS, S. M., Alakhras, J. T., & Jaganathan, P. P. (2020). Knowledge and Practices Toward Routine Medical Checkup Among Middle-Aged and Elderly People of Riyadh. *Journal of Patient Experience*, *7*(6). https://doi.org/10.1177/2374373519851003

Aram, S. A. (2021). Assessing the effect of working conditions on routine medical checkup among artisanal goldminers in Ghana. *Heliyon*, *7*(7). https://doi.org/10.1016/j.heliyon.2021.e07596

Choi, Y., An, J., Ryu, S., & Kim, J. (2022). Development and Evaluation of Machine Learning-Based High-Cost Prediction Model Using Health Check-Up Data by the National Health Insurance Service of Korea. *International Journal of Environmental Research and Public Health*, *19*(20). https://doi.org/10.3390/ijerph192013672

Costa, E., Papatsouma, I., & Markos, A. (2022). *Benchmarking distance-based partitioning methods for mixed-type data*. http://arxiv.org/abs/2203.16287

de la Selle, T., Weiss, J., & Deschanel, S. (2024). Acoustic multiplets detection based on DBSCAN and cross-correlation. *Mechanical Systems and Signal Processing*, *211*. https://doi.org/10.1016/j.ymssp.2024.111149

Dubey, A. K., Gupta, U., & Jain, S. (2022). Medical data clustering and classification using TLBO and machine learning algorithms. *Computers, Materials and Continua*, *70*(3), 4523–4543. https://doi.org/10.32604/cmc.2022.021148

Faisal, S., Sameer, S., Kamil Mohammed, I., & S Abd, M. (2021). Review of medical diagnostics via data mining techniques. *Iraqi Journal of Science*, *62*(7), 2401–2424. https://doi.org/10.24996/ijs.2021.62.7.30

Gunawan, W. (2021). Implementasi Algoritma DBScan dalam Pemngambilan Data Menggunakan Scatterplot. *Techno Xplore : Jurnal Ilmu Komputer Dan Teknologi Informasi*, *6*(2), 91–98. https://doi.org/10.36805/technoxplore.v6i2.1179

Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes. *Jurnal Teknologi Dan Sistem Komputer*, *8*(2). https://doi.org/10.14710/jtsiskom.8.2.2020.89-93

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

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*. https://doi.org/10.1016/j.ins.2022.11.139

Johari, N. A. A. M., Mohamad, N., & Isa, N. (2020). *Smart Self-Checkup for Early Disease Prediction*.

Masruriyah, A. F. N., Novita, H. Y., & Sukmawati, C. E. (2023). *Performance Evaluation of Popular Supervised Learning Algorithms Towards Cardiovascular Disease*. *8*(3), 420–426. https://doi.org/10.32493/informatika.v8i3.34103

Masruriyah, A. F. N., Novita, H. Y., Sukmawati, C. E., Arif, S. N. N., & Ramadhan, A. R. (2023). Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling. *Journal MIND Journal / ISSN*, *8*(2), 242–253. https://doi.org/10.26760/mindjournal.v8i2.242-253

Masruriyah, A. F. N., Novita, H. Y., Sukmawati, C. E., Fauzi, A., Wahiddin, D., & Handayani, H. H. (2023). Thorough Evaluation of the Effectiveness of SMOTE and ADASYN Oversampling Methods in Enhancing Supervised Learning Performance for Imbalanced Heart Disease Datasets. *International Conference on Informatics and Computing (ICIC)*.

Masruriyah, A. F. N., Sukmawati, C. E., & Dermawan, B. A. (2024). *Memahami Data Mining dengan Python: Implementasi Praktis*. https://repository.penerbiteureka.com/publications/568010/memahami-data-mining-dengan-python-implementasi-praktis

Navarro, M. S., Pelcastre-Villafuerte, B. E., Becerril-Montekio, V., & Serván-Mori, E. (2022). Overcoming the health systems' segmentation to achieve universal health coverage in Mexico. *International Journal of Health Planning and Management*, *37*(6). https://doi.org/10.1002/hpm.3538

Mia, M., Masruriyah, A. F. N., & Pratama, A. R. (2022). The Utilization of Decision Tree Algorithm In Order to Predict Heart Disease. *JURNAL SISFOTEK GLOBAL*, *12*(2), 138. https://doi.org/10.38101/sisfotek.v12i2.551

Nnoaham, K. E., & Cann, K. F. (2020). Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? *BMC Public Health*, *20*(1). https://doi.org/10.1186/s12889-020-08930-z

Penafiel, S., Baloian, N., Sanson, H., & Pino, J. A. (2021). Predicting Stroke Risk with an Interpretable Classifier. *IEEE Access*, *9*, 1154–1166. https://doi.org/10.1109/ACCESS.2020.3047195

Setiawati, E., Fernanda, U. D., Agesti, S., Iqbal, M., & Herjho, M. O. A. (2024). Implementation of K-Means, K-Medoid and DBSCAN Algorithms In Obesity Data Clustering. *IJATIS: Indonesian Journal of Applied Technology and Innovation Science*, *1*(1). https://doi.org/10.57152/ijatis.v1i1.1109

Shpigelman, E., & Shamir, R. (2023). *A feature ranking algorithm for clustering medical data*. https://doi.org/10.1101/2023.09.30.23296349

Sonjaya, C. B., Masruriyah, A. F. N., Kusumaningrum, D. S., & Pratama, A. R. (2022). The Performance Comparison of Classification Algorithm in Order to Detecting Heart Disease. *INTERNAL (Information System Journal*, *5*(2), 166–175. https://doi.org/10.32627

Sugiura, T., Takase, H., Dohi, Y., Yamashita, S., & Seo, Y. (2024). Impact of medical checkup parameters on major adverse cardiovascular events in the general Japanese population. *Preventive Medicine Reports*, *38*. https://doi.org/10.1016/j.pmedr.2024.102600

Sutramiani, N. P., Arthana, I. M. T., Lampung, P. F., Aurelia, S., Fauzi, M., & Darma, I. W. A. S. (2024). The Performance Comparison of DBSCAN and K-Means Clustering for MSMEs Grouping based on Asset Value and Turnover. *Journal of Information Systems Engineering and Business Intelligence*, *10*(1), 13–24. https://doi.org/10.20473/jisebi.10.1.13-24

Yang, W. C., Lai, J. P., Liu, Y. H., Lin, Y. L., Hou, H. P., & Pai, P. F. (2024). Using Medical Data and Clustering Techniques for a Smart Healthcare System. *Electronics (Switzerland)*, *13* (1). https://doi.org/10.3390/electronics13010140