JDMIS: Journal of Data Mining and Information Systems
Vol. 4, No. 2, 2026, p. 63~77
ISSN: ISSN: 2986-3473 | p-ISSN 2986-5271
DOI: 10.54259/jdmis.v4i2.7676

Analisis Perbandingan Metode K-Means dan DBSCAN dalam Pengelompokan Provinsi Di Indonesia Berdasarkan Kasus Penyakit

Wahyu Aditya[1], Razib Ramadhan[2], Masna Wati[3], Joan Angelina Widians[4]
[1,2,3,4]Fakultas Teknik, Universitas Mulawarman, Indonesia
Email: [1]wahyuadt054@gmail.com, [2]razhib.ramadhan06@gmail.com, [3]masnawati@fkti.unmul.ac.id, [4]angelwidians@unmul.ac.id

ABSTRAK

Penelitian ini membandingkan kinerja algoritma K-Means dan DBSCAN dalam mengelompokkan 38 provinsi di Indonesia berdasarkan enam indikator penyakit menular tahun 2025, yaitu Tuberkulosis (TBC), HIV/AIDS, kusta, malaria, dan Demam Berdarah Dengue (DBD). Ketimpangan distribusi penyakit antar wilayah menuntut adanya pemetaan yang akurat sebagai dasar intervensi kesehatan publik. Tahapan penelitian meliputi penanganan *missing value* dengan imputasi median, normalisasi Z-Score, pemodelan *clustering*, dan pengujian menggunakan metrik *Silhouette Score*, *Davies-Bouldin Index* (DBI), serta *Calinski-Harabasz Index* (CHI). Hasil evaluasi menunjukkan bahwa algoritma DBSCAN (eps=1.2, MinPts=2) secara konsisten mengungguli K-Means (K=6) pada ketiga parameter pengujian. DBSCAN memperoleh nilai *Silhouette Score* sebesar 0,6282, DBI 0,4059, dan CHI 21,8583, sedangkan K-Means hanya mencapai *Silhouette Score* 0,4395, DBI 0,6349, dan CHI 15,8653. Keunggulan DBSCAN terletak pada kemampuannya mengisolasi data yang ekstrem (*outlier*), seperti kasus malaria di Papua dan DBD di Bali, sebagai *noise*. Kesimpulannya, DBSCAN terbukti lebih tangguh dan representatif untuk pengelompokan data epidemiologis berskala nasional yang memiliki distribusi miring (*right-skewed*). Peta klaster yang dihasilkan dapat digunakan oleh pemerintah untuk memprioritaskan alokasi program kesehatan secara lebih spesifik dan tepat sasaran di tiap wilayah.

Kata Kunci: *Clustering*, DBSCAN, K-Means, Penyakit Menular, *Silhouette Score*

ABSTRACT

This study compares the performance of K-Means and DBSCAN algorithms in clustering 38 Indonesian provinces based on six infectious disease indicators in 2025, namely Tuberculosis (TB), HIV/AIDS, leprosy, malaria, and Dengue Hemorrhagic Fever (DHF). The uneven distribution of diseases across regions requires accurate mapping for targeted public health interventions. The research stages included handling missing values with median imputation, Z-Score normalization, clustering modeling, and evaluation using Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) metrics. Evaluation results indicate that the DBSCAN algorithm (eps=1.2, MinPts=2) consistently outperformed K-Means (K=6) across all three parameters. DBSCAN achieved a Silhouette Score of 0.6282, DBI of 0.4059, and CHI of 21.8583, whereas K-Means only reached a Silhouette Score of 0.4395, DBI of 0.6349, and CHI of 15.8653. DBSCAN's superiority lies in its ability to isolate extreme outliers, such as malaria cases in Papua and DHF in Bali, as noise. In conclusion, DBSCAN is proven to be more robust and representative for clustering national-scale epidemiological data with right-skewed distributions. The resulting cluster map can be utilized by the government to prioritize specific and targeted health program allocations in each region.

Keywords: *Clustering, DBSCAN, K-Means, Infectious Diseases, Silhouette Score*

Penulis Korespondensi:

Wahyu Aditya
Email: wahyuadt054@gmail.com

Article Info

Diterima: 7 Mei 2026
Direvisi: 22 Mei 2026
Disetujui: 28 Mei 2026

1. PENDAHULUAN

Penyakit menular masih menjadi masalah kesehatan masyarakat yang serius di Indonesia, baik dari segi angka kesakitan (morbiditas) maupun angka kematian (mortalitas). Tuberkulosis (TBC), HIV/AIDS, kusta, dan malaria merupakan penyakit menular yang termasuk di dalam permasalahan yang ada di Rencana Strategis Kementerian Kesehatan Tahun 2020-2024 [2]. Sementara Demam Berdarah Dengue (DBD) merupakan salah satu penyakit yang terus diperhatikan oleh pemerintah karena pernah menjadi status Kejadian Luar Biasa pada tahun 2004 [3]. Tantangan pengendalian penyakit ini semakin kompleks mengingat Indonesia merupakan negara kepulauan dengan 38 provinsi yang memiliki kondisi geografis, demografis, dan infrastruktur kesehatan yang sangat beragam. Keberagaman ini berdampak pada pola distribusi penyakit menular yang tidak merata antar wilayah. Berdasarkan data Badan Pusat Statistik Indonesia tahun 2025, terdapat enam indikator penyakit menular utama yang dipantau pada tingkat provinsi, yaitu angka penemuan dan keberhasilan pengobatan Tuberkulosis (TBC), kasus baru HIV/AIDS, angka penemuan kasus kusta per 100.000 penduduk, angka kesakitan malaria per 1.000 penduduk, serta angka kesakitan Demam Berdarah Dengue (DBD) per 100.000 penduduk. Data tersebut menunjukkan variasi yang sangat signifikan, misalnya angka kesakitan malaria di Provinsi Papua mencapai 269,44 per 1.000 penduduk, jauh melampaui rata-rata nasional sebesar 2,47, sementara kasus HIV/AIDS terkonsentrasi tinggi di Jawa Barat (9.212 kasus baru) dan Jawa Timur (10.612 kasus baru). Kondisi ini menunjukkan perlunya analisis pengelompokan provinsi secara sistematis agar kebijakan kesehatan dapat diarahkan secara tepat sasaran.

Pengelompokan atau *clustering* wilayah berdasarkan indikator penyakit menular merupakan salah satu pendekatan *data mining* yang efektif untuk mengidentifikasi pola persebaran penyakit dan mendukung pengambilan keputusan di bidang kesehatan publik [4]. Beberapa penelitian terdahulu telah membuktikan efektivitas pendekatan ini. Penelitian [5] melakukan analisis klaster dalam pengelompokan provinsi di Indonesia berdasarkan variabel penyakit menular menggunakan metode *Complete Linkage*, *Average Linkage*, dan *Ward*, dan menemukan bahwa metode hierarki dapat menghasilkan pengelompokan yang bermakna secara epidemiologis. Selain itu, penelitian [6] di Nusa Tenggara Timur menggunakan algoritma K-Means untuk mengeksplorasi pola penyakit menular di 22 kabupaten dengan variabel malaria, tuberkulosis, pneumonia, kusta, diare, demam berdarah, AIDS, dan IMS, menghasilkan tiga klaster dengan profil epidemiologis yang berbeda secara signifikan dan mendukung intervensi kesehatan yang lebih tepat sasaran. Di tingkat kabupaten/kota, penerapan K-Means untuk mengelompokkan wilayah di Provinsi Sumatera Utara berdasarkan karakteristik penyakit menular seperti TBC, kusta, malaria, dan DBD juga menunjukkan hasil yang dapat digunakan sebagai dasar perencanaan program kesehatan daerah [7].

Algoritma K-Means merupakan metode *clustering* partisi yang paling banyak digunakan karena kesederhanaan dan efisiensinya dalam mengelompokkan data. Penelitian [8] berhasil menerapkan metode K-Means untuk pengelompokan kasus COVID-19 tingkat provinsi di Indonesia dan mendapatkan akurasi sebesar 85,7%. Namun, K-Means memiliki keterbatasan mendasar, yakni hanya dapat mendeteksi klaster berbentuk bola, sensitif terhadap *outlier*, dan mengharuskan pengguna menentukan jumlah klaster (k) secara teoritis [9]. Kelemahan ini menjadi relevan karena data penyakit menular provinsi di Indonesia kerap mengandung nilai-nilai ekstrem (*outlier*), seperti yang terlihat pada data malaria di provinsi-provinsi Papua dan data kusta di Maluku Utara.

Sebagai alternatif, algoritma *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) menawarkan keunggulan berupa kemampuan mendeteksi klaster dengan bentuk sembarang *(arbitrary shapes)* dan secara otomatis mengidentifikasi *outlier* sebagai *noise* tanpa perlu menentukan jumlah klaster di awal [9][10]. Penelitian [4] membandingkan algoritma DBSCAN dan K-Means dalam pengelompokan kasus COVID-19 di tingkat global dan menemukan bahwa kedua metode menghasilkan pola klaster yang berbeda, sehingga pemilihan metode perlu disesuaikan dengan karakteristik data yang dianalisis. Selain itu, penelitian [11] juga melakukan perbandingan K-Means dan DBSCAN pada data penyebaran COVID-19 di seluruh kecamatan Provinsi Jawa Barat, dan menyimpulkan bahwa setiap metode memiliki kelebihan tersendiri bergantung pada distribusi dan dimensi data. Studi lain tentang efektivitas K-Means dan DBSCAN pada data rumah sakit menunjukkan bahwa evaluasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index* (DBI) menjadi pendekatan penting untuk menentukan metode terbaik secara objektif [10].

Penelitian [9] membandingkan performa K-Means, K-Medoids, dan DBSCAN dalam pengelompokan provinsi di Indonesia berdasarkan indikator kesejahteraan masyarakat tahun 2022, dan menemukan bahwa K-Means dan DBSCAN menghasilkan nilai *Silhouette coefficient* tertinggi (0,917) serta *Davies-Bouldin index* terendah (0,089), dengan keduanya membentuk delapan klaster yang identik. Temuan ini menegaskan bahwa perbandingan kedua metode sangat relevan dilakukan dalam konteks pengelompokan provinsi di Indonesia, terutama apabila data yang digunakan memiliki karakteristik berbeda dari data kesejahteraan, seperti data penyakit menular yang cenderung memiliki distribusi tidak normal dan mengandung *outlier* signifikan.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan performa algoritma K-Means dan DBSCAN dalam pengelompokan provinsi di Indonesia berdasarkan enam indikator kasus penyakit menular tahun 2025, yaitu angka penemuan TBC, angka keberhasilan pengobatan TBC, kasus baru HIV/AIDS, angka penemuan kasus kusta, angka kesakitan malaria, dan angka kesakitan DBD. Evaluasi perbandingan dilakukan menggunakan metrik *Silhouette Score* dan *Davies-Bouldin Index* guna menentukan metode yang paling optimal. Hasil penelitian ini diharapkan dapat memberikan rekomendasi metodologis bagi penelitian *data mining* di bidang kesehatan serta menjadi dasar penentuan prioritas intervensi program kesehatan nasional berdasarkan profil epidemiologis tiap klaster provinsi.

2. METODE PENELITIAN

Tahapan dalam penelitian ini meliputi pengumpulan data, *preprocessing*, *clustering* menggunakan algoritma K-Means dan DBSCAN, serta analisis perbandingan antara kedua algoritma tersebut. Perbandingan dievaluasi menggunakan metrik *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Adapun data yang digunakan merupakan data yang bersumber dari Badan Pusat Statistik (BPS) terkait kasus penyakit menurut provinsi pada tahun 2025. Data tersebut mencakup angka penemuan dan keberhasilan pengobatan Tuberkulosis (TBC), kasus baru HIV/AIDS, angka penemuan kasus kusta per 100.000 penduduk, angka kesakitan malaria per 1.000 penduduk, serta angka kesakitan Demam Berdarah Dengue (DBD) per 100.000 penduduk.

2.1. Preprocessing Data

Preprocessing data merupakan tahapan awal yang perlu dilakukan sebelum proses *clustering* untuk memastikan kualitas data yang digunakan memenuhi standar analisis. Adanya *missing value* dalam suatu dataset dapat menimbulkan beberapa masalah, seperti penurunan kinerja analisis, hasil yang bias, dan interpretasi yang keliru, sehingga penanganan yang tepat sangat diperlukan [12]. Pada penelitian ini, tahapan *preprocessing* terdiri dari dua langkah utama, yaitu penanganan *missing value* dan normalisasi data.

Berdasarkan hasil pemeriksaan dataset, ditemukan 10 nilai yang hilang (*missing value*) yang tersebar pada tiga variabel, yaitu variabel Kusta (6 data), Malaria (2 data), dan DBD (2 data). Metode imputasi merupakan solusi yang paling umum digunakan untuk menangani masalah dataset yang tidak lengkap, karena metode ini dapat menghindari berkurangnya jumlah dataset yang digunakan dalam proses analisis [13]. Mengingat keempat variabel yang mengandung *missing value* memiliki distribusi yang tidak normal dan condong ke kanan (*right-skewed*) dengan nilai *skewness* antara 2,40 hingga 3,72, maka imputasi dilakukan menggunakan nilai median. Imputasi *missing data* dengan metode mean mengisi data yang hilang dengan rata-rata nilai yang diketahui pada suatu variabel, namun penggunaan mean tidak disarankan pada data berdistribusi miring karena akan menghasilkan nilai imputasi yang tidak representatif akibat pengaruh nilai ekstrem [14]. Nilai median lebih tidak mudah terpengaruh oleh *outlier* dan lebih tepat digunakan pada distribusi yang miring .

Setelah penanganan *missing value*, seluruh variabel dinormalisasi menggunakan metode Z-Score Standardization. Normalisasi data penting dilakukan sebelum analisis *cluster* karena membantu menghindari dominasi atribut dengan skala besar sehingga hasil *cluster* tidak didominasi oleh satu variabel saja [15]. Hal ini menjadi penting dalam penelitian ini karena keenam variabel memiliki satuan dan rentang nilai yang sangat berbeda, misalnya variabel HIV/AIDS bernilai ratusan hingga ribuan, sementara variabel Malaria bernilai antara 0,01 hingga 269,44. Normalisasi Z-Score merupakan teknik statistik yang dapat digunakan untuk melakukan *preprocessing* serta mengembalikan skala data ke dalam rentang nilai yang seimbang antar atribut [16]. Rumus normalisasi Z-Score dapat dituliskan sebagai berikut:

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

Keterangan:
Z = Nilai hasil normalisasi
x = Nilai data
μ = Nilai rata-rata (mean)
σ = Standar deviasi

Normalisasi Z-Score dipilih karena metode ini dapat mengurangi efek *outlier* dengan mengubah data menjadi distribusi dengan rata-rata nol dan standar deviasi satu, sehingga setiap variabel memiliki kontribusi yang seimbang dalam proses *clustering*. Seluruh tahapan *preprocessing* ini diimplementasikan menggunakan bahasa pemrograman Python dengan memanfaatkan *library* Pandas, NumPy, dan Scikit-learn.

2.2. Clustering K-Means

K-Means merupakan salah satu algoritma *clustering* dengan metode partisi yang bekerja berdasarkan titik pusat (*centroid*). Metode ini membagi data ke dalam *k* kelompok sehingga data yang berkarakteristik sama dimasukkan ke dalam satu kelompok, sementara data yang berkarakteristik berbeda dimasukkan ke dalam kelompok yang lain [17]. Algoritma K-Means bertujuan meminimalkan jarak antara setiap titik data terhadap *centroid* klasternya masing-masing sehingga data dalam satu klaster memiliki kemiripan yang tinggi dan data antar klaster memiliki perbedaan yang signifikan [18].

Tahapan algoritma K-Means adalah sebagai berikut:
(1) Menentukan nilai K sebagai jumlah klaster yang diinginkan
(2) Menginisialisasi K titik centroid secara acak dari data yang tersedia
(3) Menghitung jarak setiap data terhadap seluruh centroid menggunakan euclidean distance
(4) Mengalokasikan setiap data ke klaster dengan *centroid* terdekat
(5) Memperbarui nilai *centroid* setiap klaster berdasarkan rata-rata seluruh anggota klaster
(6) Mengulangi langkah 3–5 hingga nilai *centroid* tidak berubah (*konvergen*) atau mencapai jumlah iterasi maksimum

Pengukuran jarak antar data dalam penelitian ini menggunakan Euclidean Distance dengan rumus sebagai berikut [19]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \quad (x_i - y_i)^2} \tag{2}$$

Keterangan:
$d(x,y)$ = Jarak antara data dan centroid
x_i = Nilai data ke-i
y_i = Nilai centroid ke-i
n = Jumlah variabel

Salah satu kelemahan utama K-Means adalah pengguna harus menentukan nilai *K* secara teoritis sebelum proses *clustering* dijalankan. Untuk mengatasi hal ini, penelitian ini menggunakan Elbow Method guna menentukan jumlah klaster yang optimal. Metode Elbow bekerja dengan menghitung nilai *Within-Cluster Sum of Squares* (WCSS) untuk berbagai nilai *K*, kemudian memilih nilai *K* pada titik di mana penurunan WCSS mulai melambat secara signifikan membentuk sudut seperti siku (*elbow*) pada grafik [20]. Nilai WCSS dihitung menggunakan rumus berikut:

$$WCSS \; = \; \sum_{k=1}^{K} \quad \sum_{x_i \in C_k} \quad \left|\left|x_i - \mu_k\right|\right|^2 \tag{3}$$

Keterangan:
K = Jumlah klaster
C_k = Himpunan data dalam klaster ke-k
x_i = Data ke-i
μ_k = Nilai centroid klaster ke-k

Penelitian sebelumnya menunjukkan bahwa Elbow Method merupakan pendekatan yang paling efektif dalam menentukan jumlah klaster optimal pada algoritma K-Means [21]. Oleh karena itu, penelitian ini menggunakan Elbow Method untuk menentukan nilai *K* optimal, yang kemudian dievaluasi menggunakan *Silhouette Score*, *Davies-Bouldin Index*, dan *Calinski-Harabasz Index*.

2.3. Clustering DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) merupakan algoritma *clustering* berbasis kepadatan (*density-based*) yang mengelompokkan data berdasarkan kepadatan titik data di sekitarnya. Berbeda dengan K-Means yang menggunakan metode partisi, DBSCAN tidak memerlukan jumlah klaster sebagai masukan melainkan menggunakan dua parameter utama, yaitu *epsilon* (ε) dan *MinPts* [22]. Keunggulan utama DBSCAN dibanding K-Means adalah kemampuannya mendeteksi klaster dengan bentuk sembarang (*arbitrary shapes*) serta secara otomatis mengidentifikasi data yang tidak masuk ke klaster manapun sebagai *noise* [23]. Hal ini cocok untuk penelitian ini mengingat data penyakit menular antar provinsi mengandung nilai-nilai ekstrem yang signifikan, seperti angka malaria di Papua dan Papua Pegunungan.

DBSCAN mengenal tiga jenis titik data dalam proses *clustering* [22][24]:

(1) Core Point: titik data yang memiliki jumlah tetangga di dalam radius ε lebih dari atau sama dengan nilai *MinPts*
(2) *Border Point*: titik data yang memiliki jumlah tetangga kurang dari *MinPts* namun berada di dalam radius ε dari sebuah *core point*
(3) *Noise Point*: titik data yang bukan merupakan *core point* maupun *border point* dan dianggap sebagai pencilan (*outlier*)
(4) Jika jumlah titik dalam radius ε ≥ *MinPts*, titik tersebut ditetapkan sebagai *core point* dan klaster baru dibentuk
(5) Memperluas klaster dengan memasukkan seluruh titik yang *density-reachable* dari *core point* tersebut
(6) Mengulangi proses hingga seluruh titik data telah diproses; titik yang tidak tergabung dalam klaster manapun ditandai sebagai *noise*

Tahapan algoritma DBSCAN dalam penelitian ini adalah sebagai berikut [24][25]:

(1) Menentukan nilai parameter ε dan MinPts
(2) Memilih titik data awal secara acak
(3) Menghitung jarak titik tersebut terhadap seluruh titik data menggunakan *Euclidean Distance*
(4) Jika jumlah titik dalam radius ε ≥ *MinPts*, titik tersebut ditetapkan sebagai *core point* dan klaster baru dibentuk
(5) Memperluas klaster dengan memasukkan seluruh titik yang *density-reachable* dari *core point* tersebut
(6) Mengulangi proses hingga seluruh titik data telah diproses; titik yang tidak tergabung dalam klaster manapun ditandai sebagai *noise*

Penentuan nilai parameter ε yang tepat sangat kritis karena nilai yang terlalu kecil akan menghasilkan terlalu banyak klaster kecil, sementara nilai yang terlalu besar dapat menggabungkan klaster yang berbeda menjadi satu [26]. Dalam penelitian ini, nilai ε optimal ditentukan menggunakan k-distance graph dengan menetapkan k = MinPts, yaitu dengan mengurutkan jarak setiap titik data terhadap tetangga ke-k terdekatnya secara menurun, kemudian memilih nilai ε pada titik elbow grafik tersebut [27]. Adapun nilai MinPts ditentukan berdasarkan aturan umum yaitu MinPts ≥ dimensi data + 1, sehingga dengan enam variabel dalam penelitian ini nilai MinPts yang digunakan adalah sebesar 2 hingga 5, dan nilai optimal dipilih berdasarkan hasil evaluasi metrik [29]. Penyesuaian parameter ε dan MinPts dilakukan secara iteratif untuk memperoleh pemisahan klaster yang optimal [6].

2.4. Evaluasi

Evaluasi hasil *clustering* merupakan tahapan kritis untuk mengukur kualitas pengelompokan yang dihasilkan dan menentukan metode terbaik antara algoritma K-Means dan DBSCAN secara objektif. Penelitian ini menggunakan tiga metrik evaluasi internal, yaitu *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Metrik evaluasi internal dipilih karena penilaian kualitas *clustering* dilakukan berdasarkan *dataset* dan hasil *clustering* itu sendiri, tanpa memerlukan label kebenaran eksternal (*ground truth*) [28]. Penggunaan ketiga metrik secara bersamaan bertujuan agar perbandingan antara K-Means dan DBSCAN menghasilkan kesimpulan yang lebih komprehensif dan tidak bergantung pada satu sudut pandang metrik saja [29].

a. Silhouette Score
Silhouette score memberikan ukuran seberapa baik setiap titik data ditempatkan dalam klasternya sendiri dibandingkan dengan klaster lain [30]. Nilai ini dihitung untuk setiap titik data berdasarkan dua komponen, yaitu a_i (rata-rata jarak data ke-i terhadap seluruh data lain dalam satu klaster) dan b_i (rata-rata jarak data ke-i terhadap seluruh data dari klaster lain terdekat). Rumus silhouette score untuk setiap titik data ke-i adalah sebagai berikut [31]:

$$s(i) = \frac{b_i - a_i}{max(a_i, b_i)}$$

Keterangan:
$s(i)$ = Nilai silhouette score untuk data ke-i
a_i = Rata-rata jarak data ke-i terhadap seluruh data lain dalam klaster yang sama
s_i = Rata-rata jarak data ke-i terhadap seluruh data pada klaster lain terdekat
$max(a_i, b_i)$ = Nilai maksimum antara a_i dan b_i (4)
Nilai Silhouette Score keseluruhan diperoleh dari rata-rata $s(i)$ seluruh titik data, dengan rentang nilai antara -1 hingga +1. Nilai mendekati +1 menunjukkan bahwa data berada di klaster yang tepat, nilai mendekati 0 menunjukkan data berada di perbatasan antar klaster, dan nilai negatif menunjukkan data kemungkinan salah ditempatkan [30].

b. Davies-Bouldin Index (DBI)
Davies-Bouldin Index mengevaluasi kualitas *clustering* dengan mengukur seberapa jauh jarak antar klaster relatif terhadap ukuran klaster itu sendiri [30]. Semakin kecil nilai DBI, semakin baik kualitas klaster karena menunjukkan klaster yang lebih kompak dan lebih terpisah satu sama lain. Rumus DBI adalah sebagai berikut [29]:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \quad max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \quad (5)$$

Keterangan:
k = Jumlah klaster
s_i = Rata-rata jarak titik data dalam klaster i terhadap centroid nya
s_j = Rata-rata jarak titik data dalam klaster j terhadap centroid nya
d_{ij} = Jarak antara centroid klaster i dan klaster j
Nilai DBI mendekati 0 menunjukkan kualitas *clustering* yang semakin baik [28].

c. Calinski-Harabasz Index (CHI)
Calinski-Harabasz Index, yang juga dikenal sebagai *Variance Ratio Criterion* (VRC), mengukur kualitas *clustering* berdasarkan rasio antara variasi antar klaster (*between-cluster variance*) dan variasi dalam klaster (*within-cluster variance*) [32]. Metrik ini memiliki keunggulan dibanding dua metrik sebelumnya karena bersifat *robust* terhadap data *outlier* [33]. Rumus CHI adalah sebagai berikut [32]:

$$CHI = \frac{SS_B}{SS_W} \times \frac{N - k}{k - 1} \quad (6)$$

Keterangan:
SS_B = Variasi antar klaster
SS_W = Variasi dalam klaster
N = Jumlah total data
k = Jumlah klaster
Semakin tinggi nilai CHI, semakin baik kualitas *clustering* karena menunjukkan klaster yang padat dan terpisah dengan baik [33]. Ringkasan interpretasi ketiga metrik evaluasi disajikan pada Tabel 1.

Tabel 1. Ringkasan Interpretasi Metrik Evaluasi *Clustering*

Metrik	Nilai Terbaik
Silhouette Score	Mendekati +1
Davies-Bouldin Index	Mendekati 0
Calinski-Harabasz Index	Semakin tinggi

Perbandingan antara algoritma K-Means dan DBSCAN dilakukan dengan membandingkan nilai ketiga metrik tersebut secara bersamaan. Metode yang menghasilkan *Silhouette Score* tertinggi, DBI terendah, dan CHI tertinggi secara konsisten akan

ditetapkan sebagai metode yang paling optimal untuk pengelompokan provinsi di Indonesia berdasarkan data kasus penyakit menular [28][29].

3. HASIL DAN PEMBAHASAN

3.1. Data Awal Penelitian

Data yang digunakan dalam penelitian ini diperoleh dari Badan Pusat Statistik (BPS) Indonesia yaitu data Kasus Penyakit Menurut Provinsi dan Jenis Penyakit tahun 2025. Data ini mencakup 38 provinsi sebagai objek penelitian dengan enam variabel sebagai atribut *clustering*, yaitu jumlah angka penemuan TBC , Angka Keberhasilan Pengobatan TBC, Kasus Baru HIV/AIDS, Angka Penemuan Kasus Kusta per 100.000 penduduk, Angka Kesakitan Malaria per 1.000 penduduk, dan Angka Kesakitan DBD per 100.000 penduduk. Data awal selengkapnya disajikan pada Tabel 2.

Tabel 2. Data Awal Kasus Penyakit Menurut Provinsi dan Jenis Penyakit tahun 2025

Provinsi	Jumlah Kasus Penyakit - Angka Penemuan TBC	Jumlah Kasus Penyakit - Angka Keberhasilan Pengobatan TBC	Jumlah Kasus Penyakit - HIV/AIDS Kasus Baru	Jumlah Kasus Penyakit - Penemuan Kasus Baru Kusta per 100.000 Penduduk	Jumlah Kasus Penyakit - Angka Kesakitan Malaria per 1.000 Penduduk	Jumlah Kasus Penyakit - Angka Kesakitan DBD per 100.000 Penduduk
Aceh	68	81	338	4,32	0,06	37,08
Sumatera Utara	74	79	3041	–	0,37	27,91
Sumatera Barat	62	86	578	1,11	0,02	41,63
Riau	72	86	1031	1,69	0,44	54,07
Jambi	65	86	326	1,63	0,02	37,87
Sumatera Selatan	63	89	927	3	–	45,57
Bengkulu	48	76	197	1,55	–	41,35
Lampung	65	92	1134	1,89	0,2	69,21
Kepulauan Bangka Belitung	63	85	282	3,5	0,01	104,72
Kepulauan Riau	50	82	867	1,79	0,25	91,31
DKI Jakarta	86	78	4353	4,44	0,05	95,17
Jawa Barat	97	86	9212	–	0,02	88,17
Jawa Tengah	84	86	6057	4,01	0,02	27,81
DI Yogyakarta	75	85	944	0,76	0,04	53,31
Jawa Timur	77	88	10612	5,14	0,02	49,29
Banten	112	91	2289	7,06	0,01	56,14
Bali	83	81	2024	3,01	0,01	232,87
Nusa Tenggara Barat	60	87	674	5,29	0,02	63,58
Nusa Tenggara Timur	62	85	1249	7,75	0,75	37,28
Kalimantan Barat	75	86	1113	–	0,01	36,4
Kalimantan Tengah	77	77	485	–	0,12	40,84
Kalimantan Selatan	77	78	662	3,02	0,05	13,65
Kalimantan Timur	52	83	1294	2,76	0,24	144,77
Kalimantan Utara	64	69	208	6,14	0,08	35,5
Sulawesi Utara	67	82	812	15,11	0,37	57,95
Sulawesi Tengah	84	80	732	8,65	0,41	35,61
Sulawesi Selatan	65	84	1954	8,27	0,21	45,81
Sulawesi Tenggara	72	81	600	6,69	0,23	41,03
Gorontalo	77	88	235	16,58	1,89	54,5
Sulawesi Barat	80	87	130	13,26	0,07	77,62
Maluku	63	85	734	23,02	1,67	8,07
Maluku Utara	80	75	675	56,48	0,22	22,49
Papua Barat	88	73	1099	64,97	18,63	3,4
Papua Barat Daya	82	67	529	86,89	11,53	–
Papua	58	75	2997	–	269,44	2,14

Papua Selatan	81	80	295	26,54	146,93	25,65
Papua Tengah	72	77	2057	9,88	133,67	9,34
Papua Pegunungan	41	56	551	–	36,42	–

Berdasarkan Tabel 2, teridentifikasi 10 nilai yang hilang (*missing value*) yang tersebar pada tiga variabel, yaitu variabel Kusta pada 6 provinsi (Sumatera Utara, Jawa Barat, Kalimantan Barat, Kalimantan Tengah, Papua, dan Papua Pegunungan), variabel Malaria pada 2 provinsi (Sumatera Selatan dan Bengkulu), serta variabel DBD pada 2 provinsi (Papua Barat Daya dan Papua Pegunungan). Penanganan *missing value* tersebut dilakukan pada tahap *preprocessing* yang dijelaskan pada Sub Bab 3.2.

3.2. Hasil Preprocessing Data

Tahap *preprocessing* dilakukan untuk mempersiapkan data mentah kasus penyakit menular dari 38 provinsi di Indonesia agar memenuhi asumsi dan standar kualitas yang dibutuhkan oleh algoritma *clustering*. Proses ini mencakup dua tahapan utama, yaitu penanganan data yang hilang (*missing value*) dan normalisasi data.

Berdasarkan eksplorasi data awal, teridentifikasi sebanyak 10 *missing value* yang tersebar pada tiga variabel penyakit. Rincian data yang hilang tersebut meliputi 6 data pada variabel angka penemuan kusta (tersebar di Sumatera Utara, Jawa Barat, Kalimantan Barat, Kalimantan Tengah, Papua, dan Papua Pegunungan), 2 data pada variabel angka kesakitan malaria (Sumatera Selatan dan Bengkulu), serta 2 data pada variabel angka kesakitan DBD (Papua Barat Daya dan Papua Pegunungan).

Mengingat variabel-variabel tersebut memiliki sebaran data yang tidak normal dan condong ke kanan (*right-skewed*) dengan keberadaan nilai ekstrem (*outlier*) seperti kasus malaria di wilayah Papua yang sangat tinggi, penggunaan rata-rata (*mean*) untuk mengisi kekosongan data dapat menghasilkan nilai yang bias. Oleh karena itu, penanganan *missing value* pada penelitian ini dilakukan menggunakan metode imputasi median. Nilai median dipilih karena lebih tangguh (*robust*) terhadap nilai ekstrem, sehingga representasi data setiap variabel tetap terjaga dengan baik. Secara spesifik, nilai median yang digunakan untuk proses imputasi adalah sebesar 5,21 untuk variabel kusta (mengisi kekosongan di Sumatera Utara, Jawa Barat, Kalimantan Barat, Kalimantan Tengah, Papua, dan Papua Pegunungan), 0,21 untuk variabel malaria (di Sumatera Selatan dan Bengkulu), serta 41,49 untuk variabel DBD (di Papua Barat Daya dan Papua Pegunungan).

Setelah seluruh data terisi lengkap, masalah selanjutnya adalah rentang nilai antar variabel yang memiliki perbedaan skala yang sangat besar. Sebagai contoh, variabel kasus baru HIV/AIDS memiliki rentang nilai hingga mencapai puluhan ribu kasus (seperti di Jawa Timur dan Jawa Barat), sedangkan angka kesakitan malaria didominasi oleh nilai desimal di bawah angka 1 pada banyak provinsi di wilayah barat Indonesia. Jika langsung diproses, variabel dengan skala ribuan akan mendominasi perhitungan jarak Euclidean pada algoritma K-Means maupun perhitungan kepadatan pada DBSCAN, sehingga variabel dengan skala kecil akan kehilangan pengaruhnya dalam pembentukan klaster.

Untuk menyeimbangkan bobot kontribusi setiap indikator, dilakukan transformasi data menggunakan metode *Z-Score Standardization*. Metode ini mengubah setiap titik data sehingga seluruh variabel memiliki nilai rata-rata (*mean*) sebesar nol dan standar deviasi sebesar satu. Hasil akhir dari proses normalisasi ini menghasilkan skala data desimal yang seragam. Nilai positif menunjukkan bahwa indikator penyakit provinsi tersebut berada di atas rata-rata nasional, sedangkan nilai negatif menunjukkan posisinya berada di bawah rata-rata. Cuplikan hasil normalisasi Z-Score yang telah siap digunakan untuk tahap *clustering* disajikan pada Tabel 3.

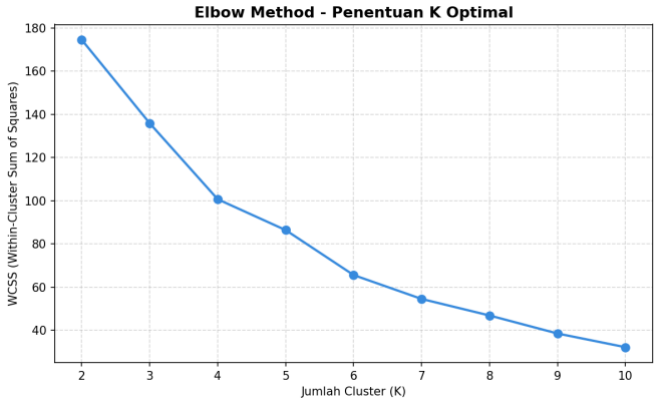
Tabel 3. Hasil Normalisasi Z-Score

Provinsi	Jumlah Kasus Penyakit - Angka Penemuan TBC	Jumlah Kasus Penyakit - Angka Keberhasilan Pengobatan TBC	Jumlah Kasus Penyakit - HIV/AIDS Kasus Baru	Jumlah Kasus Penyakit - Penemuan Kasus Baru Kusta per 100.000 Penduduk	Jumlah Kasus Penyakit - Angka Kesakitan Malaria per 1.000 Penduduk	Jumlah Kasus Penyakit - Angka Kesakitan DBD per 100.000 Penduduk
Aceh	-0,26626	-0,05265	-0,58025	-0,39491	-0,31382	-0,37111
Sumatera Utara	0,176858	-0,33845	0,601039	-0,34577	-0,30788	-0,5929
Sumatera Barat	-0,70937	0,66186	-0,47536	-0,57115	-0,31458	-0,26106
Riau	0,029152	0,66186	-0,27739	-0,53931	-0,30654	0,039825
Jambi	-0,48782	0,66186	-0,58549	-0,5426	-0,31458	-0,352
Sumatera Selatan	-0,63552	1,090565	-0,32284	-0,46739	-0,31104	-0,16576
Bengkulu	-1,74331	-0,76716	-0,64187	-0,547	-0,31104	-0,26783
Lampung	-0,48782	1,519269	-0,23237	-0,52833	-0,31113	0,406012
Kepulauan Bangka Belitung	-0,63552	0,518958	-0,60472	-0,43993	-0,31477	1,264883
Kepulauan Riau	-1,59561	0,090254	-0,34906	-0,53382	-0,31018	0,940539
DKI Jakarta	1,06309	-0,48135	1,174419	-0,38832	-0,31401	1,0339
Jawa Barat	1,87547	0,66186	3,297936	-0,34577	-0,31458	0,864593
Jawa Tengah	0,915385	0,66186	1,919114	-0,41193	-0,31458	-0,59532
DI Yogyakarta	0,250711	0,518958	-0,31541	-0,59037	-0,3142	0,021443

Jawa Timur	0,398416	0,947663	3,909774	-0,34989	-0,31458	-0,07579
Banten	2,983261	1,376368	0,272395	-0,24448	-0,31477	0,089892
Bali	0,841532	-0,05265	0,156582	-0,46684	-0,31477	4,364413
Nusa Tenggara Barat	-0,85708	0,804762	-0,4334	-0,34166	-0,31458	0,269841
Nusa Tenggara Timur	-0,70937	0,518958	-0,18211	-0,2066	-0,3006	-0,36627
Kalimantan Barat	0,250711	0,66186	-0,24155	-0,34577	-0,31477	-0,38755
Kalimantan Tengah	0,398416	-0,62425	-0,516	-0,34577	-0,31267	-0,28017
Kalimantan Selatan	0,398416	-0,48135	-0,43865	-0,46629	-0,31401	-0,9378
Kalimantan Timur	-1,4479	0,233155	-0,16245	-0,48056	-0,31037	2,233562
Kalimantan Utara	-0,56167	-1,76747	-0,63706	-0,29499	-0,31343	-0,40932
Sulawesi Utara	-0,34011	0,090254	-0,37309	0,197493	-0,30788	0,13367
Sulawesi Tengah	0,915385	-0,19555	-0,40806	-0,15718	-0,30711	-0,40666
Sulawesi Selatan	-0,48782	0,376057	0,12599	-0,17805	-0,31094	-0,15996
Sulawesi Tenggara	0,029152	-0,05265	-0,46574	-0,26479	-0,31056	-0,27557
Gorontalo	0,398416	0,947663	-0,62526	0,278201	-0,27877	0,050226
Sulawesi Barat	0,619974	0,804762	-0,67115	0,095922	-0,31362	0,609423
Maluku	-0,63552	0,518958	-0,40718	0,631778	-0,28298	-1,07276
Maluku Utara	0,619974	-0,91006	-0,43297	2,46884	-0,31075	-0,72399
Papua Barat	1,210796	-1,19586	-0,24767	2,934969	0,041849	-1,18572
Papua Barat Daya	0,767679	-2,05327	-0,49677	4,138448	-0,09414	-0,26444
Papua	-1,00479	-0,91006	0,58181	-0,34577	4,845528	-1,21619
Papua Selatan	0,693827	-0,19555	-0,59904	0,825037	2,499135	-0,64756
Papua Tengah	0,029152	-0,62425	0,171004	-0,08965	2,245171	-1,04205
Papua Pegunungan	-2,26028	-3,62519	-0,48716	-0,34577	0,382574	-0,26444

3.3. Hasil Clustering Menggunakan K-Means

Sebelum menjalankan algoritma K-Means, jumlah klaster optimal ditentukan terlebih dahulu menggunakan Elbow Method. Metode ini menghitung nilai *Within-Cluster Sum of Squares* (WCSS) untuk setiap nilai K dari 2 hingga 10, kemudian memilih nilai K pada titik di mana penurunan WCSS mulai melambat secara signifikan. Hasil grafik Elbow Method disajikan pada Gambar 1.



Gambar 1. Elbow Method Penentuan Nilai K

Berdasarkan grafik tersebut, penurunan WCSS yang cukup signifikan terjadi dari K=2 hingga K=4, setelah itu kurva mulai landai. Namun karena tidak terdapat satu titik *elbow* yang sangat tajam, dilakukan evaluasi lebih lanjut menggunakan tiga metrik evaluasi internal untuk nilai K=4, K=5, dan K=6 guna menentukan nilai K yang paling optimal secara kuantitatif.

Algoritma K-Means diimplementasikan menggunakan metode inisialisasi K-Means++ dengan jumlah pengulangan sebanyak 10 kali (*n_init*=10), di mana hasil dengan nilai *inertia* terkecil dipilih sebagai hasil akhir. Berikut hasil *clustering* untuk masing-masing nilai K yang diujikan.

Tabel 4. Hasil clustering K-Means (k = 4)

Klaster	Provinsi
Klaster 0	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah,

	Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku
Klaster 1	Papua, Papua Selatan, Papua Tengah, Papua Pegunungan
Klaster 2	Maluku Utara, Papua Barat, Papua Barat Daya
Klaster 3	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Bali

Pada K=4, terbentuk satu kelompok dominan (Klaster 0) dengan 25 provinsi yang mencakup sebagian besar wilayah Sumatera, Kalimantan, dan Sulawesi. Klaster 3 memisahkan provinsi-provinsi padat penduduk di Pulau Jawa dan Bali, sementara Klaster 1 dan Klaster 2 mengisolasi provinsi-provinsi di wilayah Papua dengan beban penyakit tropis tinggi.

Tabel 5. Hasil *clustering* K-Means (k = 5)

Klaster	Provinsi
Klaster 0	Sumatera Barat, Riau, Jambi, Sumatera Selatan, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Timur, Sulawesi Utara, Sulawesi Selatan, Gorontalo, Sulawesi Barat, Maluku
Klaster 1	Papua, Papua Selatan, Papua Tengah
Klaster 2	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Bali
Klaster 3	Maluku Utara, Papua Barat, Papua Barat Daya
Klaster 4	Aceh, Sumatera Utara, Bengkulu, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Utara, Sulawesi Tengah, Sulawesi Tenggara, Papua Pegunungan

Pada K=5, Klaster 0 yang sebelumnya beranggotakan 25 provinsi pada K=4 terpecah menjadi dua kelompok yang lebih spesifik, Klaster 0 (17 provinsi) dan Klaster 4 (9 provinsi). Papua Pegunungan yang sebelumnya berada di Klaster 1 kini bergabung ke Klaster 4 bersama beberapa provinsi lain, mengindikasikan profil epidemiologinya lebih dekat dengan kelompok tersebut dibanding provinsi Papua lainnya.

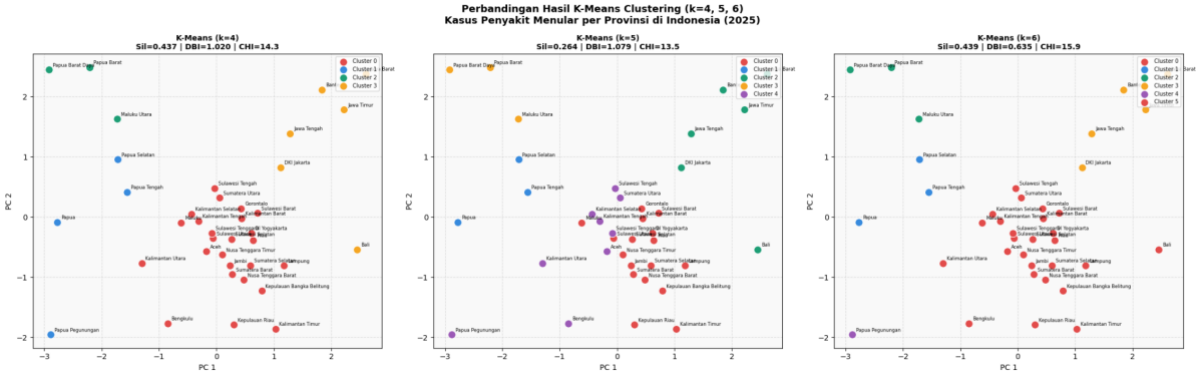
Tabel 6. Hasil clustering K-Means (k = 6)

Klaster	Provinsi
Klaster 0	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan. Bengkulu, Lampung, Kepulauan Bangka Belitung. Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku
Klaster 1	Papua, Papua Selatan, Papua Tengah
Klaster 2	Maluku Utara, Papua Barat, Papua Barat Daya
Klaster 3	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten
Klaster 4	Papua Pegunungan
Klaster 5	Bali

Pada K=6, terjadi pemisahan yang lebih granular dibandingkan K sebelumnya. Bali dipisahkan dari kelompok provinsi Jawa menjadi klaster tersendiri (Klaster 5), mencerminkan profil epidemiologis Bali yang unik dengan angka DBD tertinggi nasional sebesar 232,87 per 100.000 penduduk. Papua Pegunungan juga membentuk klaster tunggal (Klaster 4) karena karakteristik ekstremnya yang berbeda dari provinsi Papua lainnya, terutama pada variabel TBC Keberhasilan yang sangat rendah (56%) dan tidak adanya data DBD. Hasil evaluasi menggunakan ketiga metrik internal untuk seluruh nilai K yang diujikan disajikan pada Tabel 7.

Tabel 7. Perbandingan Metrik Evaluasi K-Means untuk K=4, 5, 6

K	Silhouette	DBI	CHI
4	0.4372	1.0198	14.3411
5	0.2635	1.0789	13.5304
6	0.4395	0.6349	15.8653



Gambar 2. Grafik Perbandingan K = 4, K = 5, dan K = 6

Visualisasi pada Gambar 2 menggunakan reduksi dimensi *Principal Component Analysis* (PCA) menjadi dua komponen utama (PC1 dan PC2) untuk merepresentasikan data yang semula memiliki enam variabel ke dalam ruang dua dimensi. PC1 pada sumbu horizontal merepresentasikan gradien beban penyakit menular secara keseluruhan. Provinsi dengan nilai PC1 tinggi di sisi kanan umumnya memiliki beban HIV/AIDS, kusta dan DBD tinggi seperti DKI Jakarta dan Bali, sementara provinsi

dengan nilai PC1 rendah di sisi kiri cenderung memiliki beban malaria dan kusta tinggi seperti provinsi-provinsi Papua. PC2 pada sumbu vertikal merepresentasikan intensitas penyakit tropis ekstrem terutama malaria, terlihat dari posisi Papua Selatan dan Papua Tengah yang berada jauh di atas pada sumbu ini.

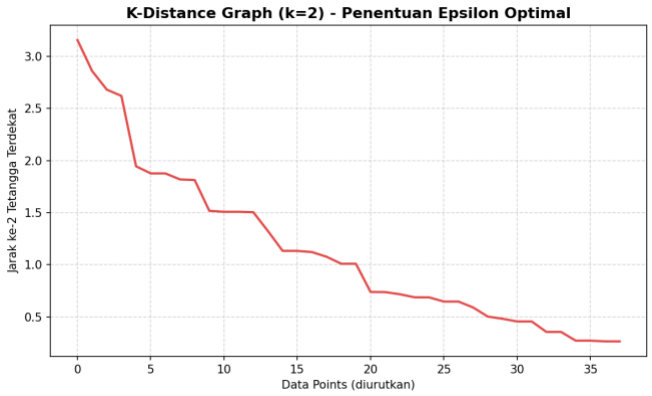
Pada grafik K=4, terlihat empat kelompok warna dengan Klaster 0 mendominasi bagian tengah mencakup 25 provinsi, klaster 1 dan 2 terdiri dari provinsi yang ada di wilayah Papua dan Maluku, sementara klaster 3 mengelompokkan provinsi di sisi kanan yang terdiri dari provinsi DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, dan Bali yang memiliki beban penyakit menular yang tinggi. Pada grafik K=5, pemisahan antar klaster terlihat kurang tegas dibandingkan K=4 maupun K=6, beberapa titik data tampak tumpang tindih yang mengkonfirmasi nilai metrik evaluasinya yang paling buruk. Pada grafik K=6, pemisahan antar klaster paling jelas, Bali (Klaster 5) dan Papua Pegunungan (Klaster 4) masing-masing berdiri sebagai titik yang sangat terisolasi di sudut grafik, mencerminkan karakteristik epidemiologis yang unik dan jauh berbeda dari provinsi lainnya.

Berdasarkan Tabel 7, nilai K=6 secara konsisten menghasilkan performa terbaik pada ketiga metrik sekaligus, yaitu *Silhouette Score* tertinggi (0,4395), DBI terendah (0,6349), dan CHI tertinggi (15,8653). Sementara itu untuk K=5 justru menunjukkan performa terburuk di semua metrik dengan *Silhouette Score* hanya 0,2635 dan CHI terendah sebesar 13,5304. K=3 dan K=4 menunjukkan hasil yang cukup baik namun tidak seunggul K=6.

Nilai *Silhouette Score* sebesar 0,4395 pada K=6 mengindikasikan bahwa data secara rata-rata ditempatkan pada klaster yang tepat dengan pemisahan antar klaster yang cukup jelas. Nilai DBI sebesar 0,6349 menunjukkan klaster yang kompak dan terpisah dengan baik, sementara CHI sebesar 15,8653 mengkonfirmasi rasio variasi antar klaster terhadap variasi dalam klaster yang paling tinggi di antara semua nilai K yang diujikan. Berdasarkan hasil evaluasi ini, K=6 ditetapkan sebagai jumlah klaster optimal untuk algoritma K-Means dalam penelitian ini.

3.4. Hasil Clustering Menggunakan DBSCAN

Penentuan nilai parameter ε pada algoritma DBSCAN dilakukan menggunakan *k-distance graph* dengan nilai *k* disesuaikan dengan nilai MinPts yang ditetapkan. Dalam penelitian ini, MinPts ditetapkan sebesar 2 dengan mempertimbangkan ukuran dataset yang kecil (n=38 provinsi) agar cluster dapat terbentuk secara optimal tanpa menghasilkan terlalu banyak *noise*. Grafik *k-distance* (k=2) disajikan pada Gambar 2.



Gambar 3. Elbow Method Penentuan Epsilon

Berdasarkan grafik tersebut, kurva turun tajam dari nilai 3,2 hingga sekitar 1,9 pada data point ke-4 hingga ke-5, kemudian mulai landai pada beberapa titik. Teridentifikasi tiga titik *elbow* yang menjadi kandidat nilai ε, yaitu pada jarak 1,8 (landai pertama di data point 4-10), 1,5 (landai kedua di data point 10-13), dan 1,2 (landai ketiga di data point 13-19). Ketiga nilai epsilon tersebut selanjutnya diujikan untuk menentukan kombinasi yang menghasilkan *clustering* paling optimal.

Algoritma DBSCAN dijalankan dengan MinPts=2 dan tiga variasi nilai epsilon sebagaimana ditentukan dari grafik *k-distance*. Berikut hasil clustering untuk masing-masing nilai epsilon.

Tabel 8. Hasil clustering DBSCAN (eps = 1.8, minpts = 2)

Klaster	Provinsi
Noise	DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Bali, Papua Barat Daya, Papua, Papua Pegunungan
Klaster 0	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku
Klaster 1	Maluku Utara, Papua Barat
Klaster 2	Papua Selatan, Papua Tengah

Pada ε=1,8, DBSCAN berhasil membentuk 3 klaster dengan 9 provinsi terdeteksi sebagai *noise*. Provinsi-provinsi yang dikategorikan sebagai *noise* merupakan wilayah dengan profil epidemiologis yang sangat ekstrem dan berbeda jauh dari provinsi lainnya. DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, dan Banten memiliki kasus HIV/AIDS yang sangat tinggi, Bali memiliki angka DBD tertinggi nasional, sementara Papua, Papua Barat Daya, dan Papua Pegunungan memiliki kombinasi angka malaria dan kusta yang sangat tinggi.

Tabel 9. Hasil clustering DBSCAN (eps = 1.5, minpts = 2)

Klaster	Provinsi
Noise	Bengkulu, DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Bali, Kalimantan Utara, Papua Barat Daya. Papua, Papua Selatan, Papua Tengah, Papua Pegunungan
Klaster 0	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku
Klaster 1	Maluku Utara, Papua Barat

Dengan ε=1,5, jumlah klaster berkurang menjadi 2 dan *noise* meningkat menjadi 13 provinsi. Bengkulu, Kalimantan Utara, Papua Selatan, dan Papua Tengah yang sebelumnya masuk klaster pada ε=1,8 kini terdeteksi sebagai *noise*, menunjukkan bahwa radius yang lebih kecil membuat syarat keanggotaan klaster semakin ketat.

Tabel 10. Hasil clustering DBSCAN (eps = 1.2, minpts = 2)

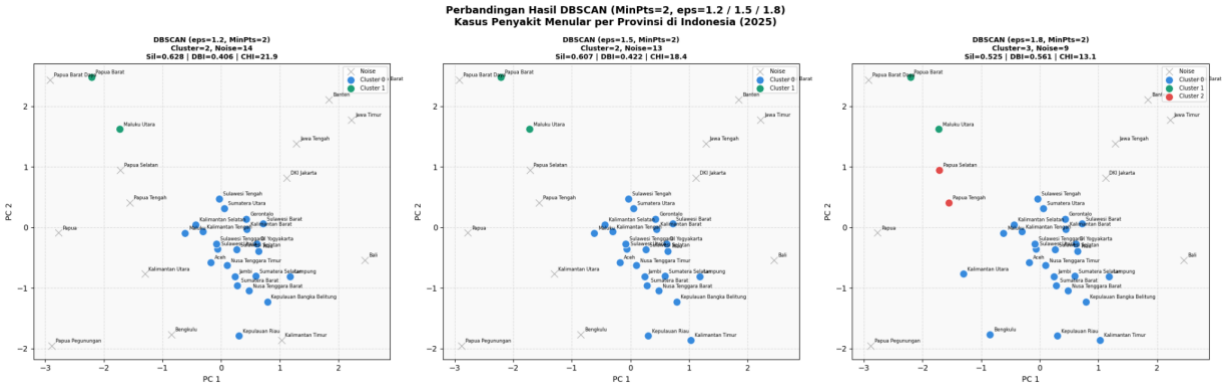
Klaster	Provinsi
Noise	Bengkulu, DKI Jakarta, Jawa Barat, Jawa Tengah, Jawa Timur, Banten, Bali, Kalimantan Timur, Kalimantan Utara, Papua Barat Daya, Papua, Papua Selatan, Papua Tengah, Papua Pegunungan
Klaster 0	Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Lampung, Kepulauan Bangka Belitung, Kepulauan Riau, DI Yogyakarta, Nusa Tenggara Barat, Nusa Tenggara Timur, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat. Maluku
Klaster 1	Maluku Utara, Papua Barat

Pada ε=1,2, terbentuk 2 klaster dengan jumlah *noise* terbanyak yaitu 14 provinsi. Kalimantan Timur kini turut terdeteksi sebagai *noise* karena memiliki angka DBD yang relatif tinggi (144,77 per 100.000 penduduk) sehingga jaraknya dalam ruang fitur melebihi radius ε=1,2 dari tetangga terdekatnya. Klaster 0 dan Klaster 1 konsisten terbentuk di ketiga variasi epsilon, Klaster 0 selalu mencakup provinsi-provinsi dengan beban penyakit sedang di Sumatera, Kalimantan, dan Sulawesi, sementara Klaster 1 selalu beranggotakan Maluku Utara dan Papua Barat yang memiliki kesamaan karakteristik beban kusta dan malaria tinggi.

Evaluasi kualitas clustering DBSCAN dilakukan menggunakan tiga metrik internal. Perhitungan metrik hanya dilakukan terhadap data yang masuk ke dalam klaster, dengan mengecualikan provinsi yang terdeteksi sebagai *noise*. Hasil evaluasi untuk ketiga nilai epsilon disajikan pada Tabel 11.

Tabel 11. Evaluasi Hasil Clustering DBSCAN

eps	Cluster	Noise	Silhouette	DBI	CHI
1.2	2	14	0.6282	0.4059	21.8583
1.5	2	13	0.6066	0.4223	18.3731
1.8	3	9	0.5248	0.5612	13.1037



Gambar 4. Grafik Perbandingan eps = 1.8, eps = 1.5, eps = 1.2

Visualisasi pada Gambar 4 menggunakan ruang PCA yang sama dengan Sub Bab 3.3, sehingga interpretasi posisi PC1 dan PC2 berlaku identik. Perbedaan utama antar ketiga grafik terletak pada jumlah titik yang dikategorikan sebagai *noise* (ditandai tanda silang abu-abu). Pada grafik ε=1,8, hanya 9 provinsi yang menjadi *noise* dan tiga klaster terbentuk, namun titik-titik dalam Klaster 0 tampak cukup tersebar di ruang PCA mengindikasikan klaster yang kurang homogen. Pada grafik ε=1,5, jumlah *noise* meningkat menjadi 13 provinsi yaitu Bengkulu, Kalimantan Utara, Papua Selatan, dan Papua Tengah yang sebelumnya masuk klaster kini bergeser menjadi *noise* karena berada di posisi yang relatif terisolasi. Pada grafik ε=1,2, titik-titik yang tersisa dalam Klaster 0 dan Klaster 1 tampak paling rapat dan terkumpul dibandingkan dua nilai epsilon sebelumnya, yang secara visual mengkonfirmasi nilai *Silhouette Score* dan DBI terbaiknya.

Berdasarkan Tabel 11, nilai ε=1,2 secara konsisten menghasilkan performa terbaik pada ketiga metrik sekaligus, yaitu *Silhouette Score* tertinggi (0,6282), DBI terendah (0,4059), dan CHI tertinggi (21,8583). Meskipun ε=1,2 menghasilkan jumlah *noise* terbanyak (14 provinsi), hal ini justru mengindikasikan bahwa DBSCAN berhasil mengidentifikasi provinsi-provinsi dengan karakteristik epidemiologis yang benar-benar berbeda dan tidak dapat dikelompokkan ke dalam klaster manapun secara valid.

Nilai *Silhouette Score* sebesar 0,6282 pada ε=1,2 tergolong cukup tinggi dan menunjukkan bahwa provinsi-provinsi yang masuk ke dalam klaster memiliki kemiripan yang kuat satu sama lain. Nilai DBI sebesar 0,4059 menunjukkan klaster yang sangat kompak dengan pemisahan antar klaster yang jelas, sementara CHI sebesar 21,8583 mengkonfirmasi rasio variansi antar klaster terhadap variansi dalam klaster yang paling tinggi. Berdasarkan hasil evaluasi ini, ε=1,2 dengan MinPts=2 ditetapkan sebagai parameter optimal untuk algoritma DBSCAN dalam penelitian ini.

3.5. Perbandingan K-Means dan DBSCAN

Perbandingan kualitas *clustering* antara algoritma K-Means (K=6) dan DBSCAN (ε=1,2, MinPts=2) dilakukan menggunakan tiga metrik evaluasi internal, yaitu *Silhouette Score*, *Davies-Bouldin Index* (DBI), dan *Calinski-Harabasz Index* (CHI). Kedua metode dibandingkan berdasarkan parameter terbaiknya masing-masing yang telah ditentukan pada Sub Bab 3.3 dan Sub Bab 3.4. Hasil perbandingan selengkapnya disajikan pada Tabel 12.

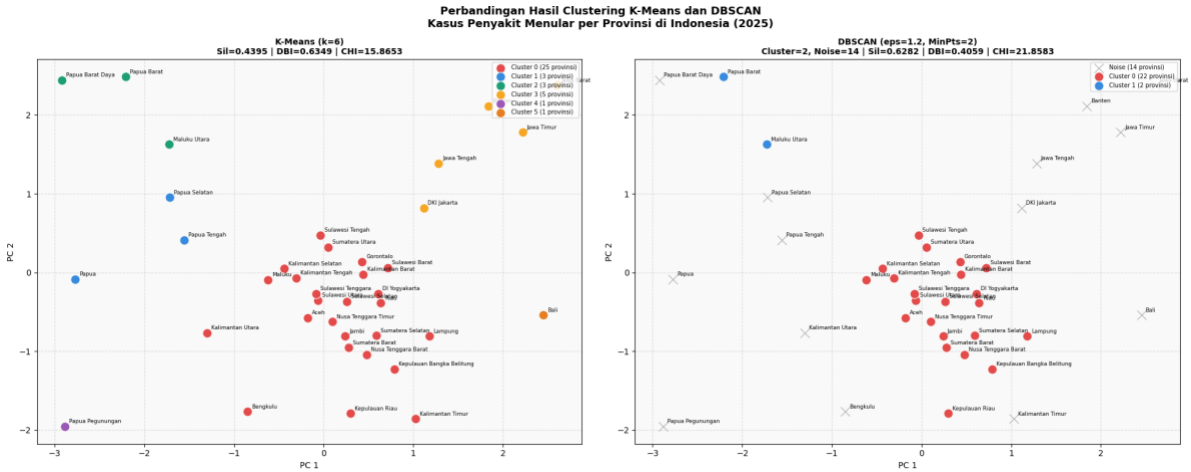
Tabel 12. Evaluasi Metrik Kinerja Algoritma K-Means dan DBSCAN

Metrik Evaluasi	K-Means (K = 6)	DBSCAN (eps = 1.2)
Silhouette Score	0.4395	0.6282
Davies-Bouldin Index	0.6349	0.4059
Calinski-Haranasz	15.8653	21.8583

Berdasarkan Tabel 12, algoritma DBSCAN dengan parameter ε=1,2 dan MinPts=2 secara konsisten unggul dibandingkan K-Means pada ketiga metrik evaluasi sekaligus. DBSCAN menghasilkan *Silhouette Score* lebih tinggi (0,6282 vs 0,4395), DBI lebih rendah (0,4059 vs 0,6349), dan CHI lebih tinggi (21,8583 vs 15,8653).

Perbedaan nilai *Silhouette Score* sebesar 0,1887 menunjukkan bahwa provinsi-provinsi yang masuk ke dalam klaster DBSCAN memiliki kemiripan yang lebih kuat satu sama lain dibandingkan klaster yang dihasilkan K-Means. Nilai DBI DBSCAN yang lebih rendah sebesar 0,2290 poin mengindikasikan klaster yang lebih kompak sekaligus lebih terpisah antar satu klaster dengan klaster lainnya. Sementara itu, selisih CHI sebesar 5,9930 poin mengkonfirmasi bahwa rasio variasi antar klaster terhadap variasi dalam klaster pada DBSCAN jauh lebih baik.

Visualisasi perbandingan hasil *clustering* kedua metode menggunakan reduksi dimensi *Principal Component Analysis* (PCA) menjadi dua komponen utama disajikan pada Gambar 5.



Gambar 5. Grafik Perbandingan K-Means (k = 6) dan DBSCAN (ε 1.2, Minpts = 2)

PC1 pada sumbu horizontal merepresentasikan gradien beban penyakit menular dari provinsi dengan beban HIV/AIDS dan DBD tinggi di sisi kanan (DKI Jakarta, Jawa Barat, Bali) hingga provinsi dengan beban malaria dan kusta tinggi di sisi kiri (Papua, Papua Barat). PC2 pada sumbu vertikal merepresentasikan intensitas penyakit tropis ekstrem terutama malaria. Pada grafik terlihat Papua Selatan dan Papua Tengah berada jauh di atas pada sumbu ini, sementara Papua Pegunungan berada di titik paling bawah kiri karena kombinasi TBC Keberhasilan terendah (56%) dengan profil variabel lain yang ekstrem.

Kedua grafik menggunakan ruang PCA yang identik sehingga posisi setiap provinsi dapat dibandingkan secara langsung. Pada grafik K-Means (kiri), seluruh titik data memiliki klaster tanpa terkecuali, Bali yang berada jauh di sisi kanan dan Papua Pegunungan yang berada di sudut paling ekstrem dipaksa masuk sebagai klaster beranggotakan satu provinsi masing-masing. Sebaliknya, pada grafik DBSCAN (kanan), provinsi-provinsi yang berada di posisi terisolasi dalam ruang PCA secara otomatis ditandai sebagai *noise* (tanda silang abu-abu), sehingga hanya Klaster 0 dan Klaster 1 yang terbentuk dengan anggota yang jauh lebih homogen dan rapat secara visual.

Secara keseluruhan, hasil perbandingan menunjukkan bahwa algoritma DBSCAN jauh lebih unggul dibandingkan K-Means dalam pengelompokan data kasus penyakit menular tingkat provinsi di Indonesia. Keunggulan utama DBSCAN terletak pada kemampuannya menangani nilai ekstrem (*outlier*) secara otomatis dengan mengisolasinya sebagai *noise*. Pada K-Means, algoritma dipaksa untuk mengalokasikan seluruh data ke dalam klaster terdekat berdasarkan jarak *Euclidean*. Hal ini menyebabkan provinsi dengan karakteristik epidemiologis ekstrem (seperti beban malaria tinggi di Papua atau DBD ekstrem di

Bali) mendistorsi titik pusat (*centroid*) dan menghasilkan klaster yang tidak representatif. Sebaliknya, pendekatan berbasis kepadatan pada DBSCAN berhasil mengabaikan anomali tersebut, sehingga klaster utama yang terbentuk menjadi jauh lebih padat, homogen, dan bermakna secara nyata. Karakteristik algoritma inilah yang membuat DBSCAN sangat cocok diterapkan pada data medis atau epidemiologis yang memiliki distribusi tidak normal (*right-skewed*), sebagaimana dibuktikan dengan perolehan nilai *Silhouette Score*, DBI, dan CHI yang secara konsisten mengungguli K-Means.

4. KESIMPULAN

Berdasarkan hasil pengujian dalam mengelompokkan 38 provinsi di Indonesia berdasarkan indikator kasus penyakit menular tahun 2025, model K-Means terbaik dicapai pada pembentukan 6 klaster (K=6). Pada titik ini, K-Means memperoleh nilai *Silhouette Score* sebesar 0,4395, *Davies-Bouldin Index* (DBI) sebesar 0,6349, dan *Calinski-Harabasz Index* (CHI) sebesar 15,8653. Sementara itu, pada algoritma DBSCAN, parameter yang paling optimal diperoleh pada konfigurasi *epsilon* (eps) = 1,2 dan MinPts = 2. Pengaturan tersebut menghasilkan evaluasi yang sangat baik dengan perolehan *Silhouette Score* sebesar 0,6282, DBI sebesar 0,4059, dan CHI sebesar 21,8583.

Berdasarkan perbandingan metrik evaluasi tersebut, algoritma DBSCAN terbukti secara signifikan lebih unggul dibandingkan K-Means. Keunggulan DBSCAN ini didasari oleh kemampuannya dalam mendeteksi dan mengisolasi nilai-nilai ekstrem (*outlier*) secara otomatis sebagai *noise*. Berbeda dengan K-Means yang memaksakan seluruh data masuk ke klaster terdekat berdasarkan jarak *Euclidean*, DBSCAN mampu mengabaikan anomali tersebut sehingga menjaga klaster utama tetap padat dan representatif. Karakteristik ini membuat algoritma DBSCAN sangat tangguh (*robust*) dan ideal untuk menangani data epidemiologis penyakit menular di Indonesia yang memiliki distribusi miring (*right-skewed*) dengan lonjakan kasus ekstrem, seperti tingginya angka malaria di Papua dan tingginya angka DBD di Bali.

Adapun hasil pemetaan menggunakan metode terbaik (DBSCAN) membagi wilayah Indonesia menjadi dua klaster utama dan satu kelompok *noise*. Pembentukan 2 klaster ini dinilai sangat relevan dan akurat dalam memotret realitas epidemiologis nasional, karena berhasil memisahkan kondisi mayoritas dengan kondisi endemi spesifik tanpa memaksakan data anomali masuk ke dalamnya. Klaster 0 beranggotakan 22 provinsi dengan tingkat beban penyakit menular yang relatif sedang dan homogen. Klaster 1 secara spesifik mengelompokkan Maluku Utara dan Papua Barat yang memiliki kesamaan beban kusta dan malaria yang sangat tinggi. Sementara itu, 14 provinsi lainnya (termasuk DKI Jakarta, Jawa Barat, Bali, dan Papua) diidentifikasi sebagai *noise* karena memiliki profil epidemiologis yang sangat ekstrem dan berbeda jauh dari provinsi lainnya. Peta pengelompokan ini dapat dijadikan landasan strategis bagi pemerintah dalam menyusun prioritas intervensi program kesehatan yang lebih presisi dan tepat sasaran.

Untuk penelitian selanjutnya, disarankan untuk mempertimbangkan penambahan variabel indikator kesehatan lainnya seperti pneumonia, diare, dan hepatitis untuk menghasilkan pengelompokan yang lebih komprehensif. Selain itu, penggunaan metode *clustering* lain seperti K-Medoids atau Gaussian Mixture Model dapat dijadikan pembanding tambahan untuk memperkuat rekomendasi metodologis, serta penerapan analisis time-series pada data beberapa tahun untuk mengidentifikasi perubahan pola distribusi penyakit antar provinsi dari waktu ke waktu. Penggunaan algoritma *clustering* berbasis kepadatan lainnya seperti *Ordering Points to Identify the Clustering Structure* (OPTICS) juga direkomendasikan, mengingat OPTICS mampu menangani dataset dengan tingkat kepadatan yang bervariasi antar wilayah, sebuah kondisi yang sangat mungkin terjadi apabila cakupan data diperluas hingga tingkat kabupaten/kota di seluruh Indonesia.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Informatika Universitas Mulawarman atas dukungan akademik, ilmu, dan pembelajaran yang diberikan selama masa perkuliahan. Ucapan terima kasih juga disampaikan kepada Badan Pusat Statistik yang telah menyediakan data sehingga penelitian ini dapat dilaksanakan dengan baik. Data yang diberikan menjadi landasan penting dalam proses penyusunan artikel ini. Penulis berharap hasil penelitian ini dapat memberikan manfaat, khususnya sebagai referensi dalam pengembangan metode *data mining* untuk memetakan serta mendukung upaya penanganan masalah kesehatan masyarakat di Indonesia.

REFERENSI

[1] K. Khairiah, Helda, dan H. N. Ghazali, "Prioritas Masalah Penyakit Menular di Kota Depok Tahun 2023," MEDIKA ALKHAIRAAT: Jurnal Penelitian Kedokteran dan Kesehatan, vol. 6, no. 2, pp. 818–828, Apr. 2025, doi: https://doi.org/10.31970/ma.v7i01.253

[2] Kementerian Kesehatan Republik Indonesia, Peraturan Menteri Kesehatan Republik Indonesia Nomor 21 Tahun 2020 tentang Rencana Strategis Kementerian Kesehatan Tahun 2020–2024, Jakarta, Indonesia, 2020. https://kesprimkom.kemkes.go.id/assets/uploads/contents/others/PMK_21_2020_Renstra_Kemenkes_Tahun_2020-2024.pdf?utm

[3] J. J. S. Cakranegara, "Upaya Pencegahan dan Pengendalian Penyakit Demam Berdarah Dengue di Indonesia Tahun 2004–2019, "Jurnal Ilmiah Indonesia, vol. 6, no. 6, pp. 2550–2561, 2021, doi: 10.36424/jpsb.v7i2.274

[4] R. Adha, N. Nurhaliza, U. Sholeha, dan Mustakim, "Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia," SITEKIN: Jurnal Sains, Teknologi dan Industri, vol. 18, no. 2, pp. 206–211, 2021. https://ejournal.uin-suska.ac.id/index.php/sitekin/article/view/12469

[5] Ulinnuha dkk., "Analisis Cluster dalam Pengelompokan Provinsi di Indonesia Berdasarkan Variabel Penyakit Menular," InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan, 2020. https://jurnal.uisu.ac.id/index.php/infotekjar/article/view/2464
[6] Aidi dkk., "Eksplorasi Pola Penyakit Menular di Nusa Tenggara Timur Menggunakan Algoritma K-Means Clustering," PREPOTIF: Jurnal Kesehatan Masyarakat, 2025. https://journal.universitaspahlawan.ac.id/index.php/prepotif/article/view/55205
[7] M. R. Syahkur dkk., "Analisis Pengelompokan Kabupaten/Kota di Provinsi Sumatera Utara Berdasarkan Penyakit Menular Menggunakan Algoritma K-Means," Informatics and Computer Engineering Journal (IJEC), 2026. https://jurnal.bsi.ac.id/index.php/ijec/article/view/11779
[8] A. N. Sihananto dkk., "Implementasi Metode K-Means untuk Pengelompokan Kasus COVID-19 Tingkat Provinsi di Indonesia," Jurnal Informatika dan Sistem Informasi, vol. 3, no. 1, pp. 76–85, 2022 https://www.researchgate.net/publication/361421458
[9] F. W. Saputri dan D. B. Arianto, "Perbandingan Performa Algoritma K-Means, K-Medoids, dan DBSCAN dalam Penggerombolan Provinsi di Indonesia Berdasarkan Indikator Kesejahteraan Masyarakat," Jurnal Teknologi Informasi, vol. 17, no. 2, 2023, doi: https://doi.org/10.47111/jti.v7i2.9558
[10] Susilo, A., "Evaluasi Efektifitas Algoritma K-Means dan DBSCAN dalam Clustering Data Rumah Sakit untuk Optimalisasi Layanan Kesehatan," Jurnal Komputer dan Informatika, 2024. https://journal.untar.ac.id/index.php/JKI/article/view/34581
[11] B. Biantara, T. Rohana, dan A. Juwita, "Perbandingan Algoritma K-Means dan DBSCAN untuk Pengelompokan Data Penyebaran Covid-19 Seluruh Kecamatan di Provinsi Jawa Barat," Scientific Student Journal for Information, Technology and Science, vol. 4, no. 1, pp. 88–94, 2023. https://santika.upnjatim.ac.id/submissions/index.php/santika/article/view/195
[12] M. R. A. Prasetya, A. M. Priyatno, dan Nurhaeni, "Penanganan Imputasi Missing Values pada Data Time Series dengan Menggunakan Metode Data Mining," Jurnal Informasi dan Teknologi, vol. 5, no. 2, pp. 52–62, 2023. https://jidt.org/jidt/article/view/324
[13] Pamuji, F. Y., Muslikh, A. R., Arief, R. M., & Muti, D., "Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil," Jurnal Informatika Polinema, 2024. https://jurnal.polinema.ac.id/index.php/jip/article/view/5031
[14] Mukarromah, S. M., "Perbandingan Imputasi Missing Data Menggunakan Metode Mean dan Metode Algoritma K-Means," BIMASTER: Buletin Ilmiah Matematika, Statistika dan Terapannya, 2015. https://jurnal.untan.ac.id/index.php/jbmstr/article/view/12425
[15] A. A. A. Daniswara dan I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," JINACS, Universitas Negeri Surabaya., 2023. https://ejournal.unesa.ac.id/index.php/jinacs/article/download/55395/43970
[16] R. G. Whendasmoro dan J. Joseph, "Analisis Penerapan Normalisasi Data Dengan Menggunakan Z-Score Pada Kinerja Algoritma K-NN," JURIKOM: Jurnal Riset Komputer, vol. 9, no. 4, pp. 872–876, 2022. https://ejurnal.stmik-budidarma.ac.id/index.php/jurikom/article/view/4526
[17] N. N. Afidah dan Masrukan, "Penerapan Metode Clustering dengan Algoritma K-Means," PRISMA: Prosiding Seminar Nasional Matematika, vol. 6, pp. 729–738, 2023. https://journal.unnes.ac.id/sju/prisma/article/download/67038/23912
[18] W. W. Pribadi, A. Yunus, dan A. S. Wiguna, "Perbandingan Metode K-Means Euclidean Distance dan Manhattan Distance pada Penentuan Zonasi Covid-19 di Kabupaten Malang," JATI: Jurnal Mahasiswa Teknik Informatika, vol. 6, no. 2, pp. 493–499, Sep. 2022. https://ejournal.itn.ac.id/index.php/jati/article/view/4808
[19] M. Iqbal, Syaripuddin, dan M. N. Huda, "Implementasi Algoritma K-Means Clustering dengan Jarak Euclidean," Jurnal Basis, Universitas Mulawarman. 2023. https://jurnal.fmipa.unmul.ac.id/index.php/Basis/article/download/1019/485
[20] B. Hartono dan V. Lusiana., "Analisis Metode Elbow SSE, Silhouette Score, dan Jaccard Stability dalam Pemilihan Jumlah Klaster Data yang Optimal," TIN: Terapan Informatika Nusantara, 2026. https://ejurnal.seminar-id.com/index.php/tin/article/view/9271
[21] P. Vania dan B. N. Sari, "Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means," Jurnal Ilmiah Wahana Pendidikan, vol. 9, no. 21, pp. 547–558, 2023, doi: https://doi.org/10.5281/zenodo.10081332
[22] S. Miftahurrahmi, Zilrahmi, N. Amalita, dan T. O. Mukhti, "DBSCAN Method in Clustering Provinces in Indonesia Based on Crime Cases in 2022," UNP Journal of Statistics and Data Science, vol. 2, no. 3, pp. 330–337, 2024, doi: 10.24036/ujsds/vol2-iss3/203. https://ujsds.ppj.unp.ac.id/index.php/ujsds/article/view/203
[23] M. R. Sani dan Darmansyah, "Penerapan Density-Based Spatial Clustering of Applications with Noise untuk Analisis Data Peserta Didik pada SMA Negeri 2 Samboja," Jurnal Teknik Informatika UNIKA Santo Thomas, 2025. https://ejournal.ust.ac.id/index.php/JTIUST/article/view/5532
[24] M. S. Hasibuan, A. H. Lubis dan M. N. Sari, "Perbandingan Algoritma Clustering DBSCAN dan K-Means," Infotech Journal, STTM Cileungsi, 2024. https://jurnal.sttmcileungsi.ac.id/index.php/infotech/article/download/1457/687
[25] P. Y. Utami, S. A. Hudjimartsu, T. A. Viona, dan H.. Sharfina, "Optimasi Parameter Algoritma DBSCAN untuk Mendeteksi Titik Panas Kebakaran Hutan dan Lahan," JEPIN: Jurnal Edukasi dan Penelitian Informatika, 2023. https://jurnal.untan.ac.id/index.php/jepin/article/view/61714
[26] H. Istiqomah, K. Nisa dan A. S. S. A, "Penerapan Algoritma DBSCAN dalam Mengidentifikasi Risiko Stroke," METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi, vol. 9, 2025, doi: https://doi.org/10.46880/jmika.Vol9No2.pp340-349
[27] K. W. Pakuani dan R. Kurniawan, "Kajian Penentuan Nilai Epsilon Optimal Pada Algoritma DMDBSCAN Dan Pemetaan Daerah Rawan Gempa Bumi Di Indonesia Tahun 2014–2020," Seminar Nasional Official Statistics, vol. 2021, no. 1, 2021. https://www.researchgate.net/publication/355840679
[28] M. F. E. Yuri, F. S. Pasaribu, A. B. Subuh, M. H. Naibaho, A. Piliang, "Evaluasi Performa Gaussian Mixture Model dan K-Means terhadap Ketidakseimbangan Data pada Clustering," Jurnal Ilmu Komputer dan Informatika, 2025. https://jurnal.globalscients.com/index.php/jiki/article/view/1221
[29] Y. Hasan, "Pengukuran Silhouette Score dan Davies-Bouldin Index pada Hasil Cluster K-Means dan DBSCAN," KAKIFIKOM, vol. 6, no. 1, pp. 60–74, 2024. https://ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/3938
[30] M. A. Pakpahan, S. Sahid, M. M. F. Simanullang, R. P. Winanda, "Analisis Hierarchical Clustering untuk Segmentasi Pelanggan pada Dataset Mall Customers," Jurnal Sistem Informasi dan Informatika, 2026. https://jurnal.unidha.ac.id/index.php/jiska/article/view/2615
[31] A. Halawa dan A. H. Lubis, "Culinary Location Clustering Using the DBSCAN Algorithm," Jurnal Informatika dan Aplikasi Sistem Informasi Engineering (JAISE), Politeknik Negeri Lhokseumawe, 2025. https://e-jurnal.pnl.ac.id/JAISE/article/view/7512

[32] M. F. Fadhillah, A. L. A. Suyoso, dan I. Puspitasari, "Customer Segmentation with Clustering Algorithm Based on Recency, Frequency, and Monetary (RFM) Attributes," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 5, no. 1, pp. 48–56, Jan. 2025, doi: 10.57152/malcom.v5v1.1491. https://journal.irpi.or.id/index.php/malcom/article/view/1491

[33] P. Aselnino dan A. W. Wijayanto, "Analisis Perbandingan Metode Hierarchical dan Non-Hierarchical dalam Pembentukan Cluster Provinsi di Indonesia Berdasarkan Indikator Women Empowerment," Jurnal Sistem Informasi dan Informatika, 2023. https://jurnal.uns.ac.id/ijas/article/download/68876/43854