

Analisa Perbandingan Algoritma K-Means dan DBSCAN Untuk Klastering Wilayah Rawan Bencana

Nathanael Kenneth Lay^{1,*}, Desi Arisandi¹, Henoch Juli Christanto²

¹ Fakultas Teknologi Informasi, Program Studi Teknik Informatika, Universitas Tarumanagara, Jakarta, Indonesia

² Fakultas Teknik, Program Studi Informatika, Universitas Kristen Indonesia, Jakarta, Indonesia

Email : ^{1,*}nathanaelkenneth.nk@gmail.com, ²desi@fti.untar.ac.id, ³henoch.christanto@gmail.com

Email Penulis Korespondensi: nathanaelkenneth.nk@gmail.com

Submitted: 17/11/2025; Accepted: 26/12/2025; Published: 26/12/2025

Abstrak—Provinsi Jawa Barat merupakan wilayah dengan tingkat kerentanan bencana yang tinggi, sehingga memerlukan pemetaan zona risiko yang akurat untuk mitigasi. Penelitian ini bertujuan melakukan analisis komparatif kinerja algoritma K-Means dan DBSCAN untuk klasterisasi wilayah rawan bencana. Perbandingan kedua algoritma ini penting karena K-Means yang berbasis sentroid sensitif terhadap outlier, sementara DBSCAN yang berbasis kepadatan secara teori lebih cocok untuk data risiko bencana yang kompleks dan mampu mengidentifikasi anomali. Pendekatan kuantitatif diterapkan pada data sekunder kejadian bencana (Banjir, Tanah Longsor, Gempa Bumi, Kebakaran, dan Cuaca Ekstrem) periode 2015-2024 di 27 kabupaten/kota. Setelah proses normalisasi menggunakan StandardScaler, kinerja model dievaluasi melalui perbandingan *Silhouette Score* (SI) dan analisis visual terhadap struktur klaster yang terbentuk. Hasil penelitian menunjukkan temuan paradoksal: meskipun K-Means (K=2) unggul secara numerik dengan SI rata-rata 0.502 dibanding DBSCAN (0.468), algoritma ini terbukti gagal secara validitas internal karena memunculkan nilai siluet negatif yang mengindikasikan kesalahan pengelompokan pada wilayah ekstrem. Sebaliknya, DBSCAN terbukti lebih superior dalam merepresentasikan struktur data alami dengan kemampuan mengisolasi 6 wilayah anomali sebagai noise. Analisis sensitivitas temporal lebih lanjut mengungkap dinamika risiko yang signifikan, di mana jumlah wilayah noise meningkat drastis dari 5 wilayah pada periode 2015-2019 menjadi 10 wilayah pada periode 2020-2024. Hal ini mengindikasikan bahwa pola bencana di Jawa Barat menjadi semakin tidak teratur dan sulit diprediksi, sehingga DBSCAN direkomendasikan sebagai metode yang lebih robust untuk pemetaan risiko yang kompleks.

Kata Kunci: Clustering; DBSCAN; Jawa Barat; K-Means; Silhouette Score; Pemetaan Risiko Bencana

Abstract—West Java Province exhibits high disaster vulnerability, necessitating accurate risk zone mapping for mitigation purposes. This study aims to conduct a comparative performance analysis of K-Means and DBSCAN algorithms for clustering disaster-prone areas. The comparison is important because K-Means, as a centroid-based algorithm, is sensitive to outliers, whereas DBSCAN, as a density-based method, is theoretically more suitable for complex disaster-risk data and capable of identifying anomalies. A quantitative approach was applied to secondary data on disaster incidents (Floods, Landslides, Earthquakes, Fires, and Extreme Weather) for the 2015-2024 period across 27 regencies/cities. Following normalization using StandardScaler, model performance was evaluated through Silhouette Score (SI) comparison and visual analysis of the formed cluster structures. The results reveal a paradoxical finding: although K-Means (K=2) numerically outperformed DBSCAN (0.468) with an average SI of 0.502, it demonstrated internal validity failure due to negative silhouette scores indicating misclassification in extreme regions. Conversely, DBSCAN proved superior in representing the natural data structure with its capability to isolate 6 anomalous regions as noise. Further temporal sensitivity analysis revealed significant risk dynamics, where the number of noise regions increased drastically from 5 regions in the 2015-2019 period to 10 regions in the 2020-2024 period. This indicates that disaster patterns in West Java are becoming increasingly irregular and unpredictable, positioning DBSCAN as the recommended robust method for complex risk mapping.

Keywords: Clustering; DBSCAN; West Java; K-Means; Silhouette Score; Disaster Risk Mapping

1. PENDAHULUAN

Bencana alam, baik yang bersifat geologis maupun hidrometeorologis, merupakan isu kritis yang menuntut respons terencana dan berbasis data, terutama di wilayah dengan kerentanan tinggi seperti Indonesia [1]. Provinsi Jawa Barat secara geografis dan demografis merupakan salah satu wilayah dengan tingkat kerentanan bencana alam tertinggi di Indonesia. Berdasarkan data historis, frekuensi kejadian bencana hidrometeorologi seperti banjir dan tanah longsor, serta bencana geologi seperti gempa bumi, menunjukkan tren yang signifikan dan terus meningkat di berbagai kabupaten/kota [2]. Distribusi kejadian bencana ini tidak terjadi secara seragam, melainkan membentuk pola spasial yang kompleks, mengikuti karakteristik topografi yang bervariasi mulai dari pesisir hingga pegunungan, serta dipengaruhi oleh faktor lingkungan yang dinamis. Dalam manajemen risiko bencana, efektivitas upaya mitigasi sangat bergantung pada pemahaman mendalam mengenai pola spasial tersebut [3]. Pengambilan kebijakan yang berbasis bukti (evidence-based policy) memerlukan adanya pemetaan zona kerawanan yang akurat sebagai instrumen strategis untuk alokasi sumber daya, perencanaan tata ruang, dan program kesiapsiagaan masyarakat yang terfokus. Tanpa pemetaan yang presisi, intervensi pemerintah berisiko menjadi tidak tepat sasaran dan tidak efisien. Oleh karena itu, diperlukan adopsi metodologi data mining seperti klasterisasi untuk mengidentifikasi pola risiko yang tersembunyi dan membagi wilayah ke dalam kelompok kerentanan yang homogen.

Masalah mendasar yang dihadapi dalam pemetaan risiko [4] saat ini adalah karakteristik data kebencanaan yang sangat heterogen dan mengandung anomali (outlier). Data kejadian bencana alam seringkali memiliki distribusi yang tidak normal, di mana sebagian besar wilayah mungkin memiliki kejadian rendah, namun terdapat beberapa wilayah dengan frekuensi kejadian yang sangat ekstrem. Pendekatan konvensional menggunakan teknik data mining

atau pengelompokan (clustering) telah banyak diadopsi untuk memecahkan masalah ini. Namun, mayoritas penelitian dan penerapan praktis masih sangat bergantung pada algoritma K-Means. Meskipun K-Means populer karena efisiensi komputasinya, algoritma ini memiliki limitasi fundamental yang krusial: ia bekerja dengan asumsi bahwa kluster harus berbentuk sferis (bulat) dan memiliki sensitivitas yang sangat tinggi terhadap outlier. Algoritma ini akan memaksakan setiap titik data, seberapa pun ekstremnya, untuk masuk ke dalam salah satu kluster rata-rata. Akibatnya, wilayah dengan karakteristik bencana yang unik atau ekstrem seringkali "tertelan" ke dalam kluster umum, sehingga menyembunyikan tingkat urgensi risiko yang sebenarnya. Hal ini berpotensi menghasilkan bias informasi yang berbahaya bagi strategi mitigasi bencana.

Sebagai solusi untuk mengatasi kelemahan metode berbasis partisi tersebut, penelitian ini mengajukan penerapan algoritma berbasis kepadatan, yaitu Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Secara teoritis, DBSCAN menawarkan pendekatan yang berbeda secara fundamental karena tidak membatasi bentuk kluster harus sferis dan dirancang khusus untuk memisahkan data yang memiliki kepadatan rendah sebagai noise atau outlier. Kemampuan untuk mengisolasi noise ini dipandang sangat relevan dengan data kebencanaan yang seringkali memiliki anomali ekstrem. Oleh karena itu, solusi yang diharapkan dari penelitian ini adalah ditemukannya metode klusterisasi yang paling representatif dan mewakili terhadap struktur data asli, melalui analisis komparatif yang ketat antara algoritma K-Means (sebagai baseline) dan DBSCAN (sebagai penantang).

Untuk mendukung posisi penelitian ini dalam khazanah keilmuan, telah dilakukan penelusuran terhadap berbagai penelitian sejenis dalam kurun waktu lima tahun terakhir. Algoritma K-means diterapkan untuk mengelompokkan daerah rawan bencana di 34 provinsi di Indonesia [5]. Studi ini menggunakan data kejadian tanah longsor tahun 2021 dan berhasil memetakan Indonesia ke dalam dua kluster, di mana Jawa Barat teridentifikasi sebagai daerah rawan. Penelitian ini mengonfirmasi bahwa K-Means mumpuni sebagai metode standar untuk analisis makro, namun tidak mengeksplorasi metode lain untuk menangani variasi data yang lebih detail. Secara lebih spesifik K-Means pernah digunakan untuk memprediksi daerah rawan bencana di 27 kabupaten/kota di Jawa Barat periode 2012-2021 [6]. Dengan menggunakan lima variabel bencana, mereka berhasil membentuk tiga kluster (rendah, sedang, tinggi). Studi ini memberikan bukti konsep (proof of concept) yang valid mengenai penggunaan clustering di Jawa Barat, namun tetap mewarisi kelemahan K-Means yang tidak memperhitungkan deteksi outlier secara eksplisit.

Menyadari kelemahan K-Means, Herviany, dkk. [7] mencoba melakukan komparasi dalam keluarga algoritma partisi, yaitu antara K-Means dan K-Medoids untuk kasus tanah longsor di Jawa Barat. Hasil evaluasi mereka menggunakan Davies-Bouldin Index (DBI) menunjukkan bahwa K-Medoids memberikan hasil yang sedikit lebih baik karena lebih robust terhadap outlier dibandingkan K-Means. Temuan ini secara empiris membuktikan bahwa algoritma partisi standar (K-Means) bukanlah satu-satunya pilihan terbaik dan membuka peluang untuk eksplorasi algoritma non-partisi. DBSCAN diterapkan untuk pengelompokan daerah rawan bencana gempa bumi di Pulau Sumatera [8]. Hasil penelitian ini sangat signifikan karena berhasil membuktikan kapabilitas DBSCAN dalam mengidentifikasi 190 titik atau poin sebagai noise atau outlier. Hal ini menunjukkan bahwa DBSCAN mampu menangkap pola bencana yang unik yang tidak dapat dilakukan oleh metode partisi biasa. Temuan serupa juga didapatkan ketika DBSCAN digunakan untuk klusterisasi daerah rawan banjir di Kabupaten Deli Serdang, yang memperkuat argumen bahwa metode berbasis kepadatan sangat cocok untuk data spasial yang tidak beraturan [9].

Meskipun demikian, superioritas algoritma sangat bergantung pada data. Saat K-Means dibandingkan dengan DBSCAN pada data COVID-19 global justru menemukan bahwa K-Means lebih unggul secara metrik Silhouette Index dibandingkan DBSCAN [10]. Hal ini menunjukkan bahwa tidak ada algoritma yang selalu menang, dan performa sangat bergantung pada karakteristik dataset spesifik yang dianalisis.

Berdasarkan telaah literatur tersebut, ditemukan kesenjangan penelitian (Research Gap) yang nyata. Meskipun penerapan K-Means di Jawa Barat sudah banyak dilakukan [6], dan komparasi antar metode partisi sudah ada [7], serta penerapan DBSCAN terbukti sukses di wilayah lain [11], belum ada studi komprehensif yang secara spesifik membenturkan kinerja K-Means melawan DBSCAN menggunakan dataset multi-hazard 5 jenis bencana, yaitu gempa bumi, tanah longsor, kebakaran, cuaca ekstrem, dan banjir pada tingkat kabupaten/kota di Provinsi Jawa Barat dengan rentang data jangka panjang (10 tahun). Kekurangan studi ini menyebabkan belum adanya rekomendasi metodologis yang pasti mengenai algoritma mana yang paling tepat untuk menangani kompleksitas data bencana di wilayah ini.

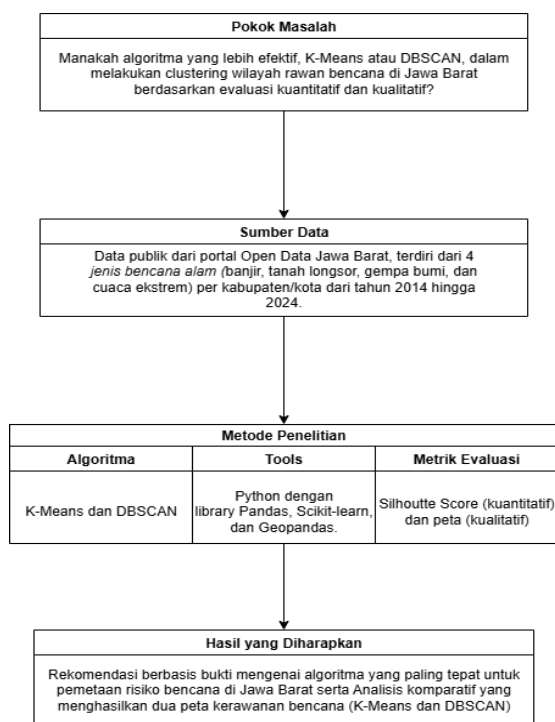
Oleh karena itu, penelitian ini dirancang untuk mengisi kesenjangan literatur dengan melakukan analisis komparatif mendalam terhadap kinerja algoritma K-Means dan Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Fokus utama studi ini adalah klusterisasi wilayah rawan bencana di Provinsi Jawa Barat berdasarkan data historis periode 2015–2024. Pemilihan kedua algoritma ini didasarkan pada perbedaan fundamental mekanismenya: K-Means yang berbasis centroid dan DBSCAN yang berbasis kepadatan, yang diharapkan dapat memberikan perspektif berbeda dalam memetakan pola spasial bencana. Tujuan utamanya adalah mengevaluasi efektivitas kedua algoritma tersebut melalui dua pendekatan komplementer. Secara kuantitatif, performa akan diukur menggunakan parameter Silhouette Score untuk menilai kohesi dan separasi antar kluster. Secara kualitatif, penelitian ini akan membedah kemampuan masing-masing model dalam merepresentasikan kompleksitas data lapangan, khususnya sejauh mana algoritma tersebut mampu menangani outlier atau pencilan data yang sering kali merepresentasikan kejadian bencana ekstrem namun langka. Luaran yang ditargetkan dari penelitian ini adalah sebuah rekomendasi berbasis bukti empiris mengenai metode klusterisasi yang paling superior untuk karakteristik data bencana di Jawa Barat. Hasil analisis ini diharapkan tidak hanya menjadi kontribusi teoretis, tetapi juga instrumen praktis bagi Badan Penanggulangan Bencana Daerah (BPBD) dan pemangku kepentingan terkait. Dengan tersedianya

peta risiko yang lebih akurat dan presisi, strategi mitigasi dapat diarahkan secara lebih spesifik pada wilayah dengan karakteristik risiko ekstrem. Pada akhirnya, hasil komparasi ini akan berfungsi sebagai panduan strategis berbasis data (data-driven policy) dalam mengoptimalkan alokasi sumber daya dan merumuskan kebijakan mitigasi yang lebih fokus, efisien, dan tepat sasaran di tingkat provinsi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain studi komparatif (*comparative study*) untuk mengevaluasi kinerja algoritma K-Means dan DBSCAN. Kerangka kerja penelitian diadopsi dari standar industri Cross-Industry Standard Process for Data Mining (CRISP-DM) [12][12] yang disederhanakan menjadi empat tahapan utama yang sistematis, sebagaimana diilustrasikan pada Gambar 1.



Gambar 1. Tahapan Penelitian

Gambar 1 menyajikan tahapan penelitian yang dimulai dari tahapan pertama yaitu penentuan pokok masalah atau pondasi dalam penelitian ini, dilanjutkan dengan pengumpulan sumber data dari sumber yang relevan, setelah itu masuk ke tahapan penelitian yang menggunakan 2 algoritma utama yaitu K-Means dan DBSCAN dibantu dengan menggunakan bahasa pemrograman python sampai pengujian metrik evaluasi dan diakhiri dengan penarikan kesimpulan dari hasil penelitian yang didapatkan.

2.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh melalui studi dokumentasi pada portal resmi Open Data Provinsi Jawa Barat. Dataset mencakup rekapitulasi historis kejadian bencana alam selama periode 10 tahun, terhitung mulai 1 Januari 2015 hingga 31 Desember 2024. Populasi dan unit analisis penelitian meliputi seluruh wilayah administrasi di Provinsi Jawa Barat yang terdiri dari 18 kabupaten dan 9 kota (total 27 kabupaten/kota). Variabel fitur yang diekstraksi difokuskan pada jumlah kejadian (frequency) untuk lima jenis bencana utama, yaitu: (1) Banjir, (2) Tanah Longsor, (3) Gempa Bumi, (4) Kebakaran, dan (5) Cuaca Ekstrem.

Data mentah yang dikumpulkan awalnya terpisah dalam format CSV. Selanjutnya, dilakukan proses pra-pemrosesan berupa penyaringan (filtering) tahun dan penggabungan (merging) dataset. Data kemudian diagregasi dari tingkat kejadian per desa/kecamatan menjadi total kejadian per kabupaten/kota, menghasilkan matriks fitur final berdimensi 27 x 5 yang siap untuk tahap pemrosesan selanjutnya.

2.3 Preprocessing Data

Tahap pra-pemrosesan data merupakan langkah fundamental dalam alur kerja *data mining* untuk memastikan kualitas input sebelum masuk ke tahap pemodelan. Data mentah kejadian bencana seringkali memiliki format yang tidak konsisten, skala yang timpang, dan struktur yang belum siap olah. Dalam penelitian ini, seluruh proses pra-pemrosesan

diimplementasikan menggunakan bahasa pemrograman Python dengan bantuan pustaka Pandas untuk manipulasi data tabular dan *Scikit-learn* untuk transformasi fitur [13]. Rangkaian proses ini terdiri dari tiga tahapan utama.

2.3.1 Integrasi dan Agregasi Data

Data awal diperoleh dalam bentuk lima berkas terpisah (comma-separated values) yang masing-masing merepresentasikan jenis bencana berbeda. Langkah pertama adalah melakukan penyaringan (filtering) untuk memastikan data yang diambil hanya berada dalam rentang waktu studi, yaitu tahun 2015 hingga 2024. Selanjutnya, dilakukan proses integrasi penggabungan kelima dataset tersebut. Karena data asli tercatat pada tingkat administratif desa atau kecamatan, dilakukan operasi agregasi spasial (grouping) dengan menjumlahkan total frekuensi kejadian bencana untuk setiap kabupaten/kota. Proses ini mereduksi granularitas data menjadi 27 entitas wilayah unik sesuai dengan jumlah kabupaten/kota di Jawa Barat.

2.3.2 Transformasi Matriks Fitur

Setelah diagregasi, struktur data ditransformasi menggunakan teknik *pivoting* untuk membentuk matriks fitur final. Matriks ini berdimensi 27 x 5, di mana baris merepresentasikan 27 kabupaten/kota dan kolom merepresentasikan 5 variabel fitur bencana (Banjir, Tanah Longsor, Gempa Bumi, Kebakaran, dan Cuaca Ekstrem). Struktur ini adalah format standar yang dibutuhkan oleh algoritma clustering untuk melakukan perhitungan jarak antar-objek.

2.3.3 Normalisasi Data

Analisis statistik deskriptif awal menunjukkan adanya variansi yang sangat ekstrem antar variabel. Variabel 'Cuaca Ekstrem' dan 'Tanah Longsor' memiliki rentang nilai yang sangat besar (hingga ribuan kejadian), sementara variabel 'Gempa Bumi' memiliki rentang nilai yang jauh lebih kecil (puluhan kejadian). Dalam algoritma berbasis jarak seperti K-Means yang menggunakan *Euclidean Distance*, fitur dengan skala nilai yang besar akan mendominasi perhitungan jarak, sehingga fitur dengan skala kecil akan terabaikan dan dianggap tidak signifikan [14].

Untuk mengatasi bias skala tersebut, penelitian ini menerapkan teknik normalisasi *Standard Scaler (Z-Score)* [15]. Teknik ini mentransformasi distribusi data pada setiap fitur sehingga memiliki rata-rata (μ) nol dan standar deviasi (S) satu. Persamaan matematis transformasi Z-score dinyatakan sebagai berikut [16]:

$$z = \frac{(x-\mu)}{s} \quad (1)$$

Dimana z adalah nilai data baru yang telah dinormalisasi, selanjutnya x adalah nilai data asli, μ adalah Nilai rata-rata (mean) dari kolom fitur tempat data x berada, dan S adalah Nilai standar deviasi sampel dari kolom fitur tempat data x berada. Untuk mendapatkan standar deviasi sampel (S) rumusnya adalah sebagai berikut [17]:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \quad (2)$$

Dimana S adalah Simbol untuk standar deviasi sampel, ini adalah hasil akhir yang dicari, \sum adalah simbol sigma atau penjumlahan, ini berarti hasil perhitungan yang ada di kanannya akan dijumlahkan semua dari data pertama ($i=1$) hingga data terakhir (n), X_i mewakili setiap nilai data individual dalam sampel data, \bar{x} mewakili nilai rata-rata (mean) dari seluruh data dalam satu kolom dan n adalah jumlah total baris data dalam satu kolom, dalam kasus ini $n = 27$ (karena ada 27 kabupaten / kota) [18]. Data hasil normalisasi inilah yang menjamin bahwa setiap jenis bencana memberikan kontribusi bobot yang setara dalam pembentukan kluster, baik pada K-Means maupun DBSCAN.

2.4 Pemodelan Clustering

Pada tahap ini, matriks data yang telah dinormalisasi diproses menggunakan dua algoritma *clustering* dengan pendekatan berbeda untuk membandingkan efektivitasnya dalam memetakan wilayah rawan bencana.

2.4.1 Klastering K-Means

K-Means adalah metode berbasis partisi yang bertujuan membagi N objek data ke dalam K kluster yang saling lepas [19]. Prinsip utama algoritma ini adalah meminimalkan variansi intra kluster, di mana setiap objek dialokasikan ke kluster dengan centroid (titik pusat) terdekat. Proses alokasi ini menggunakan perhitungan Jarak Euclidean (Euclidean Distance) sebagai ukuran kesamaan antar objek [20].

Jika dimisalkan terdapat dua kabupaten, A dan B, dengan fitur-fitur bencana f_1, f_2, f_3, f_4 , maka jarak antara keduanya ($d_{A,B}$) dihitung menggunakan persamaan [21]:

$$d(A, B) = \sqrt{(A_{f_1} - B_{f_1})^2 + (A_{f_2} - B_{f_2})^2 + (A_{f_3} - B_{f_3})^2 + (A_{f_4} - B_{f_4})^2} \quad (3)$$

Dimana $d(A, B)$ adalah jarak antara kabupaten A dan B, A_{f_1} adalah Nilai fitur 1 (banjir) untuk Kabupaten A, B_{f_1} adalah Nilai fitur 1 (banjir) untuk Kabupaten B, A_{f_2} adalah Nilai fitur 2 (tanah_longsor) untuk Kabupaten A, B_{f_2} adalah Nilai fitur 2 (tanah_longsor) untuk Kabupaten B, A_{f_3} adalah Nilai fitur 3 (gempa_bumi) untuk Kabupaten A, B_{f_3} adalah Nilai fitur 3 (gempa_bumi) untuk Kabupaten B, A_{f_4} adalah Nilai fitur 4 (cuaca_ekstrem) untuk Kabupaten

A, dan B_{f_4} adalah Nilai fitur 4 (cuaca_ekstrem) untuk Kabupaten B [22]. Proses algoritma berjalan secara iteratif dengan langkah-langkah sebagai berikut:

- a. Inisialisasi: Menentukan jumlah kluster (K) berdasarkan hasil evaluasi Elbow Method, lalu memilih K titik centroid awal secara acak.
- b. Alokasi (Assignment): Menghitung jarak setiap titik data ke semua centroid menggunakan rumus Euclidean Distance, lalu memasukkan titik data tersebut ke dalam kluster dengan centroid terdekat.
- c. Pembaruan (Update): Menghitung ulang posisi centroid baru dengan mencari nilai rata-rata (mean) dari seluruh titik data yang menjadi anggota kluster tersebut.
- d. Iterasi: Mengulangi langkah alokasi dan pembaruan hingga posisi centroid konvergen (tidak berubah lagi).

2.4.2 Klustering DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) diterapkan untuk mengatasi kelemahan K-Means dalam menangani outlier. Algoritma ini mengelompokkan titik data yang memiliki kepadatan tinggi dan memisahkannya dari titik data yang berada di area berkepadatan rendah [23].

Berbeda dengan K-Means, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) tidak memerlukan penentuan jumlah kluster di awal. Algoritma ini mendefinisikan kluster sebagai area dengan densitas titik data yang tinggi yang dipisahkan oleh area dengan densitas rendah. Algoritma ini bekerja berdasarkan dua parameter utama [24]:

- a. Epsilon (ϵ): Radius lingkungan maksimum di sekitar sebuah titik data.
- b. MinPts: Jumlah minimum titik data yang harus ada dalam radius epsilon (termasuk titik itu sendiri) untuk membentuk wilayah padat.

Berdasarkan kedua parameter tersebut, DBSCAN mengklasifikasikan setiap titik data dalam tiga kategori [25]:

- a. *Core Point* (Titik Inti): Titik yang memiliki setidaknya sejumlah MinPts titik tetangga dalam radius epsilon. Titik ini berada di pusat kluster.
- b. *Border Point* (Titik Batas): Titik yang berada dalam jangkauan radius epsilon dari Core Point, namun memiliki tetangga kurang dari MinPts. Titik ini menjadi bagian tepi dari kluster.
- c. *Noise* (Derau/Pencilan): Titik yang bukan Core Point dan tidak terjangkau oleh Core Point manapun. Titik ini dianggap sebagai anomali atau outlier.

Dalam penelitian ini, penentuan nilai optimal untuk epsilon dan MinPts dilakukan melalui eksperimen grid search sistematis guna menemukan konfigurasi yang paling akurat dalam mengisolasi noise (wilayah bencana ekstrem) tanpa memecah kluster utama secara berlebihan.

2.5 Evaluasi Kinerja Model

Untuk memastikan objektivitas dalam membandingkan kinerja algoritma K-Means dan DBSCAN, penelitian ini menggunakan pendekatan evaluasi ganda yang mencakup aspek kuantitatif dan kualitatif.

2.5.1 Evaluasi Kuantitatif (*Silhouette Score*)

Validasi internal model dilakukan menggunakan metrik Silhouette Score (SI). Metrik ini dipilih karena kemampuannya mengukur kualitas kluster tanpa memerlukan label kebenaran dasar (ground truth), yang sesuai dengan karakteristik unsupervised learning. Silhouette Score mengukur seberapa dekat setiap titik data dengan kluster tempatnya berada (kohesi) dibandingkan dengan kluster tetangga terdekatnya (separasi) [26].

Nilai siluet untuk sebuah titik data i dihitung menggunakan Persamaan:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

Dimana $a(i)$ adalah rata-rata jarak titik i ke semua titik lain dalam kluster yang sama (kohesi), $b(i)$ adalah Rata-rata jarak titik i ke semua titik di kluster tetangga terdekat (separasi), dan $\max\{a(i), b(i)\}$ artinya nilai terbesar antara $a(i)$ dan $b(i)$. Mengapa menggunakan $\max(a, b)$? Untuk menormalkan skor agar tetap berada pada rentang -1 sampai 1. Karena jika jarak ke kluster lain $b(i)$ lebih besar dari jarak ke kluster sendiri $a(i)$, maka $b(i) - a(i)$ akan positif, serta Jika jarak ke kluster sendiri lebih besar (buruk), maka $b(i) - a(i)$ akan negatif. Dengan membagi selisih $b(i) - a(i)$ dengan nilai terbesar antara a dan b , perbandingan menjadi relatif terhadap skenario terburuk yang mungkin [27].

Nilai rata-rata SI dari seluruh sampel digunakan sebagai indikator performa global model. Rentang nilai SI berkisar antara -1 hingga +1 dengan interpretasi sebagai berikut [28]:

- a. Nilai mendekati +1: Menunjukkan struktur kluster yang sangat baik (*strong*), di mana objek berada jauh dari kluster tetangga.
- b. Nilai mendekati 0: Menunjukkan kluster yang tumpang tindih (*overlapping*) atau berada di perbatasan keputusan.
- c. Nilai negatif: Mengindikasikan kesalahan pengelompokan (*misclassification*), di mana objek ditempatkan pada kluster yang salah

2.5.2 Evaluasi Kualitatif (analisis representasi)

Selain skor numerik, evaluasi kualitatif dilakukan melalui analisis visual terhadap peta hasil klusterisasi dan *scatter plot*. Fokus utama evaluasi ini adalah untuk menilai kemampuan algoritma dalam merepresentasikan struktur alami

data kebencanaan yang kompleks. Secara spesifik, analisis ini meninjau bagaimana algoritma menangani *outlier* atau anomali:

- Apakah algoritma memaksakan *outlier* masuk ke dalam kluster umum (seperti kecenderungan *K-Means*)?
- Atau apakah algoritma mampu mengisolasi *outlier* tersebut sebagai *noise* (seperti fitur DBSCAN)?

3. HASIL DAN PEMBAHASAN

3.1 Analisis Karakteristik Data

Proses penelitian ini diawali dengan tahap pengumpulan dan persiapan data yang komprehensif, sebuah fondasi krusial sebelum melangkah ke tahap pemodelan (*modeling*). Sesuai dengan kerangka kerja metodologis yang telah dirancang, data yang digunakan merupakan data sekunder yang bersumber dari portal resmi Open Data Provinsi Jawa Barat. Data mentah yang berhasil dihimpun mencakup rekam jejak historis kejadian bencana selama satu dekade penuh, terhitung mulai 1 Januari 2015 hingga 31 Desember 2024, yang meliputi 27 wilayah administratif kabupaten/kota di Jawa Barat. Fokus penelitian dipusatkan pada lima variabel fitur utama, yaitu kejadian Banjir, Tanah Longsor, Gempa Bumi, Kebakaran, dan Cuaca Ekstrem. Melalui proses merging dan agregasi spasial, data mentah tersebut ditransformasi menjadi matriks fitur final berdimensi 27×5 , yang merepresentasikan total akumulasi kejadian per wilayah.

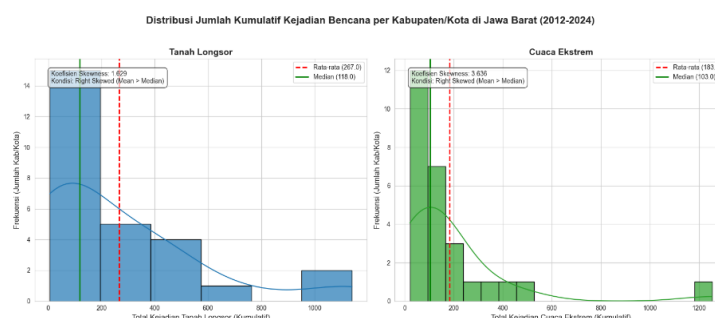
Untuk memperoleh pemahaman mendalam (*data understanding*) mengenai karakteristik dataset sebelum dilakukan pemrosesan lebih lanjut, dilakukan analisis statistik deskriptif. Analisis ini bertujuan untuk memetakan pola sebaran, ukuran pemusatan, dan tingkat variasi data, yang menjadi basis argumentasi bagi langkah pra-pemrosesan selanjutnya. Hasil rekapitulasi statistik deskriptif dari data aktual disajikan secara rinci pada Tabel 1.

Tabel 1. Statistik Deskriptif Data Kejadian Bencana (2015–2024)

Variabel	Rerata (Mean)	Variansi (Var)	Standar Deviasi (Std)	Min	Max
Banjir	107.4	7039.2	83.9	5	261
Tanah Longsor	289.8	63252.3	251.5	6	1073
Gempa Bumi	21.8	723.6	26.9	0	104
Cuaca Ekstrem	232.0	79411.2	281.8	20	1223
Kebakaran	114.6	4761.0	69.0	12	529

Berdasarkan Tabel 1, dapat ditarik sejumlah observasi analitis yang fundamental terkait struktur data kebencanaan di Jawa Barat. Temuan yang paling mencolok adalah adanya disparitas atau kesenjangan skala nilai yang sangat ekstrem antar variabel. Sebagai ilustrasi, variabel 'Gempa Bumi' memiliki rentang nilai (*range*) yang relatif sempit, mulai dari 0 hingga maksimum 104 kejadian, dengan nilai rata-rata hanya 21,8. Sebaliknya, variabel 'Cuaca Ekstrem' menunjukkan skala yang jauh lebih masif dengan rentang nilai antara 20 hingga 1.223 kejadian, serta rata-rata mencapai 232,0. Perbedaan skala ini menjadi semakin signifikan ketika meninjau nilai Variansi (*Var*). Variabel 'Cuaca Ekstrem' mencatatkan variansi sebesar 79.411,2 dan 'Tanah Longsor' sebesar 63.252,3. Tingginya nilai variansi ini mengindikasikan bahwa sebaran data pada kedua jenis bencana tersebut sangat lebar, heterogen, dan berfluktuasi tajam antar wilayah. Sebagai perbandingan, variansi pada variabel 'Gempa Bumi' jauh lebih rendah (723,6), yang menandakan sebaran data yang lebih homogen. Implikasi dari temuan statistik ini sangat krusial terhadap pemilihan strategi pemodelan. Algoritma clustering yang berbasis jarak, khususnya *K-Means* yang mengandalkan perhitungan Euclidean Distance, diketahui memiliki sensitivitas yang sangat tinggi terhadap perbedaan skala variabel. Apabila data mentah ini digunakan secara langsung tanpa transformasi, variabel dengan magnitudo nilai absolut dan variansi terbesar dalam hal ini 'Cuaca Ekstrem' dan 'Tanah Longsor' akan memberikan "bobot" yang tidak proporsional dan mendominasi perhitungan jarak antar objek. Akibatnya, kontribusi dari variabel berskala kecil seperti 'Gempa Bumi' akan menjadi terabaikan (*dwarfed*), sehingga berpotensi menghasilkan struktur kluster yang bias dan tidak merepresentasikan profil risiko bencana yang sesungguhnya.

Selain disparitas skala, karakteristik distribusi data juga diperiksa secara visual untuk mendeteksi keberadaan anomali. Gambar 2 menyajikan visualisasi histogram distribusi frekuensi kejadian untuk dua variabel dengan variansi tertinggi.



Gambar 2. Histogram Sebaran Data (Aktual) untuk Variabel Tanah Longsor dan Cuaca Ekstrem

Sebagaimana diperlihatkan pada Gambar 2, bentuk distribusi data untuk variabel 'Tanah Longsor' dan 'Cuaca Ekstrem' sangat menjulur ke kanan atau right-skewed. Secara interpretatif, pola distribusi ini mengungkapkan fenomena ketimpangan wilayah: mayoritas kabupaten/kota (sekitar 20–22 wilayah) memiliki frekuensi kejadian bencana yang relatif rendah hingga sedang, namun terdapat segelintir kecil wilayah (sekitar 1–2 wilayah) yang mencatatkan frekuensi kejadian sangat ekstrem, mencapai angka maksimum 1.073 dan 1.223 kejadian. Titik-titik data ekstrem yang terpisah jauh dari mayoritas populasi inilah yang dalam analisis statistik dikenal sebagai outlier atau pencilan. Kehadiran outlier yang terkonfirmasi melalui nilai maksimum dan standar deviasi yang tinggi ini menjadi tantangan metodologis yang serius. Algoritma K-Means, yang bekerja dengan meminimalkan variansi rata-rata ke pusat kluster (centroid), cenderung rentan terdistorsi oleh kehadiran outlier, yang dapat menarik posisi centroid menjauh dari pusat massa kluster yang sebenarnya. Sebaliknya, kondisi data yang "kotor" dengan noise ini secara teoritis menjadi domain yang dapat ditangani lebih baik oleh algoritma berbasis kepadatan seperti DBSCAN.

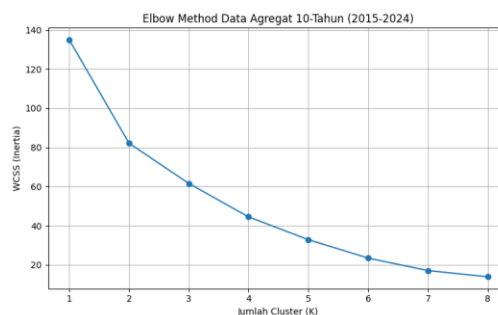
Keseluruhan temuan dari analisis deskriptif ini mulai dari perbedaan skala antar variabel yang ekstrem hingga konfirmasi keberadaan outlier yang signifikan melalui distribusi right-skewed memberikan justifikasi teknis yang kuat dan tak terbantahkan mengenai perlunya tahap pra-pemrosesan lanjutan. Oleh karena itu, sesuai dengan rancangan penelitian, dataset ini wajib melalui proses normalisasi menggunakan teknik StandardScaler (transformasi Z-score). Langkah ini memastikan bahwa setiap variabel bencana ditransformasi ke dalam skala yang setara, sehingga memberikan kontribusi bobot yang adil dan seimbang dalam proses pembentukan kluster pada tahap pemodelan selanjutnya.

3.2 Penentuan Parameter Optimal

Sebelum melakukan komparasi kinerja, langkah fundamental yang harus dilakukan adalah menentukan konfigurasi parameter terbaik untuk masing-masing algoritma. Hal ini bertujuan untuk memastikan bahwa perbandingan dilakukan secara adil pada performa puncak masing-masing model.

3.2.1 Parameter K-Means (Jumlah Kluster)

Penentuan jumlah kluster (K) optimal untuk algoritma K-Means dilakukan menggunakan pendekatan visual Elbow Method. Metode ini menghitung nilai inersia atau Within-Cluster Sum of Squares (WCSS) untuk setiap variasi K mulai dari 1 hingga 8.



Gambar 3. Grafik Elbow Method pada Data Agregat 10 Tahun (2015–2024)

Berdasarkan Gambar 3, terlihat penurunan inersia yang drastis terjadi pada transisi dari $K=1$ ke $K=2$. Penurunan masih berlanjut hingga $K=3$, namun setelah titik tersebut, kurva mulai melandai (*leveling off*). Secara visual, "siku" terbentuk di sekitar $K=3$.

Namun, Elbow Method terkadang bersifat subjektif. Oleh karena itu, verifikasi dilakukan menggunakan metrik objektif Silhouette Score. Berdasarkan pengujian eksperimental (lihat Tabel 2), K-Means dengan $K=2$ menghasilkan skor siluet rata-rata tertinggi sebesar 0.502, jauh mengungguli konfigurasi $K=3$ yang hanya mencapai skor 0.389. Hal ini mengindikasikan bahwa struktur data bencana Jawa Barat secara alami terpartisi lebih baik ke dalam dua kelompok utama dibandingkan tiga kelompok. Oleh karena itu, model $K=2$ dipilih untuk analisis selanjutnya 4.

Tabel 2. Peringkat Kinerja K-Means Berdasarkan Silhouette Score

Rank	K	Silhouette Score
1	2	0.5020560938827064
2	3	0.38977128831920593
3	4	0.37308425631469216
4	6	0.33974098018207866
5	7	0.33484039881275734
6	5	0.33138982890566676
7	9	0.30318938789641225
8	8	0.29068526161598535
9	10	0.25303880725935496



Berdasarkan Tabel 2, temuan menarik diperoleh. Model K-Means dengan K=2 menghasilkan *Silhouette Score* tertinggi (SI = 0.502). Secara signifikan, model dengan K=3 (yang secara intuitif sering dipilih untuk kategori "Rendah, Sedang, Tinggi") justru menghasilkan skor yang jauh lebih rendah (SI = 0.389). Hal ini dapat terjadi berkaitan dengan karakteristik *elbow method* dan *silhouette score* yang berbeda:

- a. Elbow method
 1. Metode ini hanya mengukur WCSS/Inertia, yaitu seberapa padat kluster (seberapa dekat titik ke pusat kluster-nya).
 2. Metode ini tidak mempertimbangkan seberapa baik kluster terpisah satu sama lain.
 3. Penentuan "siku" (*elbow*) seringkali subjektif. Seseorang bisa melihat K=3, orang lain mungkin K=4.
- b. Sedangkan, silhouette score
 1. Mengukur berdasarkan cohesion (kepadatan): seberapa dekat titik dengan kluster nya sendiri.
 2. Mengukur berdasarkan Separation (pemisahan): seberapa jauh titik dari kluster tetangganya sendiri.
 3. Metode *silhouette score* ini juga memberikan skor objektif yang menyeimbangkan kedua faktor tersebut.

Hal ini memberikan indikasi awal bahwa data kebencanaan Jawa Barat secara alami tidak terpartisi dengan baik ke dalam 3 kelompok, melainkan lebih condong ke 2 kelompok. Oleh karena itu, untuk pengujian hipotesis, model K-Means terbaik yang akan digunakan adalah dengan parameter K=2.

3.2.2 Parameter DBSCAN (Eps dan MinPts)

Berbeda dengan *K-Means*, tantangan pada DBSCAN adalah menyeimbangkan antara kualitas kluster (densitas) dan jumlah data yang terklasifikasi sebagai *noise*. Karena parameter *Epsilon* (ϵ) dan *Minimum Points* (*MinPts*) saling mempengaruhi, dilakukan eksperimen sistematis (*grid search*) untuk mengevaluasi berbagai kombinasi parameter berdasarkan *Silhouette Score*. Hasil pengujian parameter DBSCAN disajikan pada Tabel 3.

Tabel 3. Peringkat Kinerja DBSCAN Berdasarkan Silhouette Score

eps	min samples	Silhouette Score	n clusters	noises
0.5000	3	0.6755	2	19
1.1000	2	0.5785	2	9
0.7000	3	0.5177	2	15
0.7000	2	0.5177	2	15
0.9000	2	0.4702	3	10
1.5000	2	0.4680	2	6
1.3000	3	0.4680	2	6
1.3000	2	0.4680	2	6
1.5000	3	0.4680	2	6

Berdasarkan Tabel 3, terlihat adanya *trade-off* yang signifikan antara nilai *Silhouette Score* dan retensi data. Model dengan skor tertinggi (Rank 1, SI = 0.676) secara matematis sangat solid, namun mengklasifikasikan 19 dari 27 wilayah (70%) sebagai *noise*. Model ini dinilai terlalu agresif dan tidak praktis untuk tujuan pemetaan karena gagal memberikan label pada mayoritas wilayah studi.

Sebagai solusi komparatif yang seimbang, penelitian ini memilih konfigurasi Model Rank 6 ($\epsilon=1.5$, *MinPts*=2). Meskipun memiliki skor siluet yang lebih rendah (0.468) dibandingkan Rank 1 atau Rank 2, model ini menghasilkan jumlah *noise* sebanyak 6 wilayah. Jumlah ini identik dengan jumlah anggota "Kluster Ekstrem" yang dihasilkan oleh *K-Means* (K=2). Pemilihan ini strategis untuk memungkinkan perbandingan *apple-to-apple*: meneliti bagaimana kedua algoritma memperlakukan 6 wilayah anomali yang sama apakah dipaksa masuk kluster (oleh *K-Means*) atau diisolasi (oleh DBSCAN).

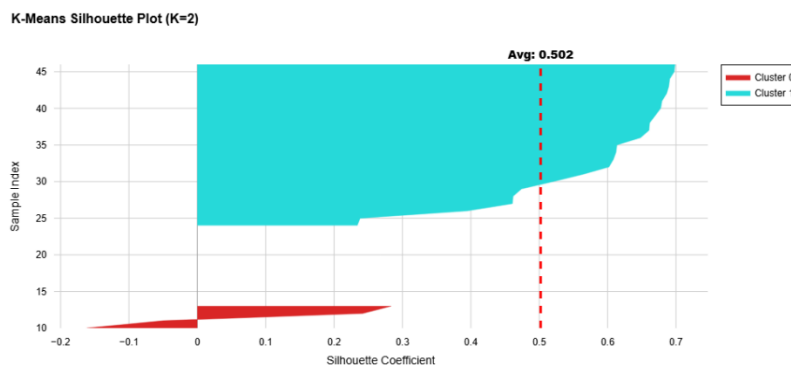
3.3 Komparasi Kinerja Algoritma

Setelah parameter optimal ditetapkan, dilakukan perbandingan langsung antara kinerja model K-Means (K=2) dan DBSCAN (epsilon=1.5, MinPts=2). Evaluasi tidak hanya didasarkan pada skor rata-rata global, tetapi juga membedah validitas internal kluster untuk melihat bagaimana setiap algoritma menangani anggota kluster secara individual. Ringkasan perbandingan kinerja kedua algoritma disajikan pada Tabel 4.

Tabel 4. Perbandingan Kinerja K-Means dan DBSCAN

Metrik	Algoritma K-Means	Algoritma DBSCAN
Parameter	K = 2	Eps = 1.5, min_samples = 2
Jumlah Kluster	2	2
Jumlah Noise	0	6
SI Rata-rata	0.502	0.468
Interpretasi Kualitas Analisis Internal	Struktur Lemah Gagal (Struktur 2 kluster tidak akurat dan tercemar)	Struktur Lemah Berhasil (Mengisolasi 6 outlier + menemukan 1 kluster unik)

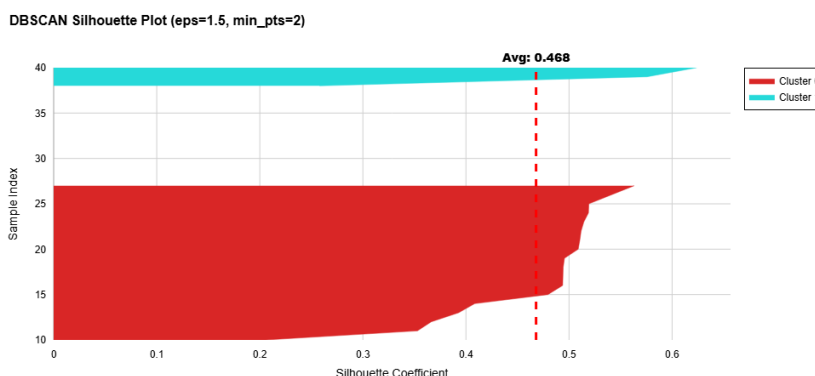
Secara kuantitatif, K-Means tampak lebih unggul dengan rata-rata *Silhouette Score* sebesar 0.502, dibandingkan DBSCAN yang mencatatkan skor 0.468. Namun, analisis mendalam terhadap distribusi skor siluet per individu (validitas internal) menyingkap kelemahan fundamental pada hasil *K-Means*. Evaluasi internal pada kluster *K-Means* menunjukkan adanya inkonsistensi. Meskipun skor rata-ratanya tinggi, terdapat anggota kluster khususnya pada kelompok wilayah ekstrem yang memiliki nilai siluet negatif seperti yang terlihat pada Gambar 4. Secara spesifik, Kabupaten Bandung (-0.049) dan Kota Bogor (-0.161) ditempatkan pada kluster yang salah (*misclassified*). Hal ini terjadi karena mekanisme *centroid* pada *K-Means* memaksakan wilayah-wilayah tersebut untuk bergabung ke dalam kluster terdekat meskipun secara jarak *Euclidean* mereka sebenarnya jauh dari pusat kluster tersebut. Fenomena ini mengonfirmasi bahwa tingginya skor rata-rata *K-Means* bersifat semu, karena menutupi kegagalan algoritma dalam menangani variasi data lokal.



Gambar 4. Silhouette Plot K-Means (K=2)

Dari Gambar 4 silhoutte plot K-Means tersebut dapat terlihat bahwa ada beberapa data (kabupaten/kota) yang memiliki *silhoutte score* bernilai negatif atau di bawah 0, namun hal ini berhasil ditutupi oleh daerah dengan *silhoutte score* yang jauh lebih tinggi daripada *silhoutte* rata-rata.

Sebaliknya, DBSCAN menunjukkan kinerja yang lebih stabil dan representatif seperti terlihat pada visual di Gambar 5. Meskipun skor rata-ratanya sedikit lebih rendah, DBSCAN tidak memiliki anggota kluster dengan nilai siluet negatif. Algoritma ini secara cerdas mengisolasi 6 wilayah termasuk Kabupaten Bandung dan Kota Bogor yang gagal dikelompokkan oleh *K-Means* sebagai *noise* atau anomali. Dalam konteks kebencanaan, keenam wilayah *noise* ini bukanlah data yang salah, melainkan merepresentasikan wilayah dengan karakteristik bencana "ekstrem" (frekuensi kejadian sangat tinggi) yang memang berbeda secara signifikan dari pola umum wilayah lain di Jawa Barat.



Gambar 5. Silhouette Plot DBSCAN

Dengan demikian, meskipun *K-Means* unggul secara numerik, DBSCAN terbukti lebih superior secara kualitatif. Kemampuan DBSCAN mengidentifikasi *noise* memberikan peta risiko yang lebih akurat, menghindari bias penyederhanaan yang dilakukan oleh *K-Means*.

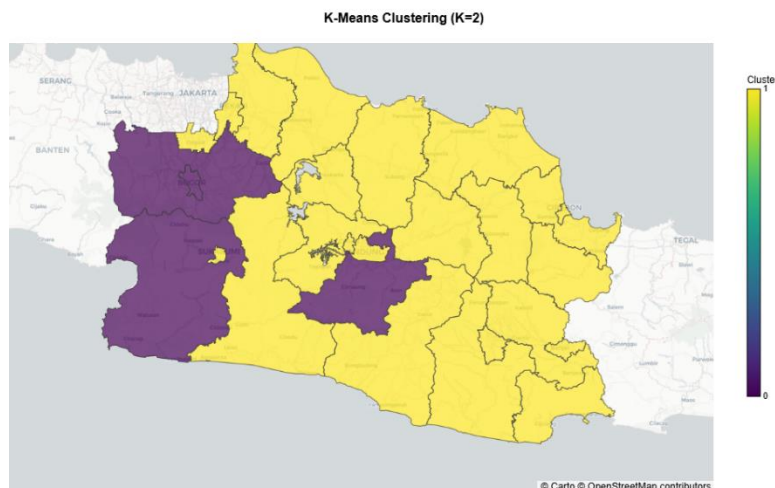
3.4 Visualisasi Peta

Untuk memahami sebaran geografis risiko bencana, hasil pengelompokan divisualisasikan ke dalam peta tematik. Visualisasi ini memperjelas perbedaan mendasar antara cara kerja *K-Means* dan DBSCAN dalam memetakan wilayah.

3.4.1 Peta Risiko Model K-Means

Hasil klasterisasi *K-Means* (K=2) membagi Jawa Barat menjadi dua zona risiko yang kontras. Gambar 6 memperlihatkan dominasi Kluster 1 (Kuning) yang mencakup mayoritas wilayah administrasi di Jawa Barat.

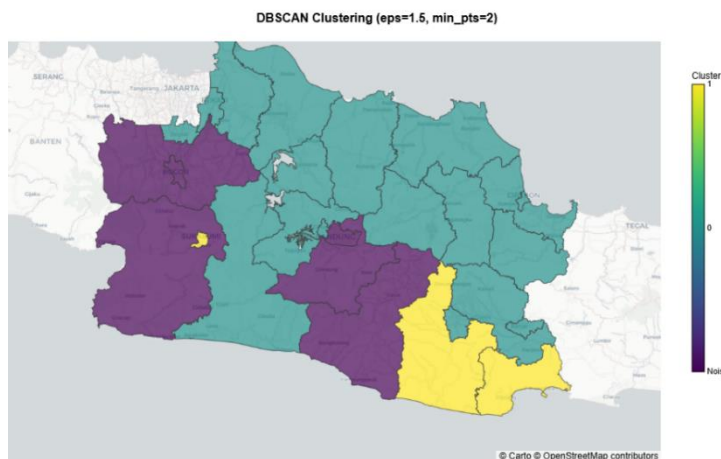
Sebaliknya, Klaster 0 (Ungu) tampak terpusat hanya pada beberapa wilayah spesifik yang meliputi Kabupaten Bogor, Kota Bogor, Kabupaten Sukabumi, dan wilayah Bandung.



Gambar 6. Peta Sebaran Klaster Bencana Menggunakan K-Means (K=2)

3.4.2 Peta Risiko Model DBSCAN

Hasil clustering DBSCAN berhasil mengelompokan daerah di Jawa Barat menjadi 3 jenis seperti pada visualisasi di Gambar 7 menampilkan struktur risiko yang lebih granular dan realistis. Algoritma ini membagi wilayah menjadi Klaster 0 (Biru - Risiko Menengah/Normal), Klaster 1 (Kuning - Risiko Spesifik Gempa), dan *Noise* (Ungu Gelap - Risiko Ekstrem).



Gambar 7. Peta Sebaran Klaster Bencana Menggunakan DBSCAN

3.5 Analisis Sensitivitas Temporal

Untuk menguji konsistensi performa algoritma terhadap perubahan tren waktu, dilakukan analisis sensitivitas dengan membagi data ke dalam dua periode lima tahunan: Periode I (2015–2019) dan Periode II (2020–2024).

Hasil pengujian menunjukkan dinamika yang signifikan. Algoritma *K-Means* menunjukkan sifat statis dengan konsisten menghasilkan $K=2$ sebagai model terbaik di kedua periode, yang mengindikasikan ketidakmampuan algoritma ini dalam merespons perubahan kompleksitas data. Sebaliknya, DBSCAN menunjukkan adaptabilitas tinggi. Pada Periode I, DBSCAN mendeteksi 5 wilayah sebagai *noise*. Namun, pada Periode II, jumlah wilayah yang terdeteksi sebagai *noise* meningkat dua kali lipat menjadi 10 wilayah.

Peningkatan jumlah *noise* ini secara implisit mengungkap bahwa pola kejadian bencana di Jawa Barat dalam lima tahun terakhir telah bergeser menjadi semakin ekstrem, acak, dan sulit diprediksi (*unpredictable*). Kemampuan DBSCAN untuk menangkap fenomena peningkatan entropi data ini menegaskan keunggulannya sebagai alat pemetaan risiko yang lebih sensitif terhadap realitas perubahan iklim dan lingkungan dibandingkan *K-Means* yang cenderung menyederhanakan masalah.

3.6 Pembahasan dan Interpretasi Hasil Penelitian

Untuk *K-Means*, sebagai algoritma berbasis partisi, memiliki kelemahan fundamental, yaitu asumsi bentuk kluster sferikal dan keharusan mengalokasikan setiap data point ke dalam kluster. Dalam data kejadian bencana Jawa Barat

yang dicirikan oleh distribusi right-skewed dan adanya anomali risiko ekstrem, K-Means gagal merepresentasikan struktur data secara akurat. Kegagalan ini dibuktikan dengan munculnya skor siluet negatif pada wilayah ekstrem (Kabupaten Bandung dan Kota Bogor), yang secara matematis mengindikasikan bahwa titik data tersebut salah diklasterkan (misclassified) dan jaraknya lebih dekat ke klaster tetangga.

Sebaliknya, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), sebagai algoritma berbasis kepadatan, terbukti lebih unggul dalam merepresentasikan struktur data yang kompleks. DBSCAN dirancang untuk mengidentifikasi klaster dengan bentuk arbitrer dan secara inheren mampu mengisolasi outlier sebagai noise. Dalam model optimalnya, DBSCAN berhasil mengisolasi 6 wilayah sebagai noise, yang secara kontekstual diinterpretasikan sebagai zona risiko bencana ekstrem yang unik, sehingga model ini lebih robust dan representatif untuk tujuan pemetaan risiko.

Temuan ini konsisten dengan literatur yang membahas kelemahan K-Means dan keunggulan DBSCAN dalam data yang tersebar dan non-linear. Penelitian terdahulu juga telah membuktikan kapabilitas DBSCAN dalam menangani outlier dan menemukan klaster yang tidak beraturan pada data kebencanaan di Pulau Sumatera dan klasterisasi daerah rawan banjir di kota Fuzhou di China [5][11]. Lebih lanjut, analisis sensitivitas temporal menguatkan keunggulan adaptif DBSCAN. Model DBSCAN menunjukkan bahwa jumlah wilayah noise meningkat drastis dari 5 menjadi 10 pada periode 2020–2024. Peningkatan ini mengindikasikan bahwa pola bencana di Jawa Barat menjadi semakin tidak stabil dan sulit diprediksi. Kemampuan DBSCAN untuk menyesuaikan parameter dan mendeteksi dinamika anomali ini menjadikannya metode yang lebih valid dan tangguh untuk analisis risiko spasial-temporal yang kompleks.

4. KESIMPULAN

Karakteristik fundamental data kebencanaan di Provinsi Jawa Barat, menunjukkan tingkat heterogenitas dan sebaran spasial yang tinggi dan tidak seragam, merupakan tujuan utama dari perbandingan mendalam antara algoritma berbasis partisi K-Means dan algoritma berbasis kepadatan DBSCAN, menggunakan data kejadian bencana periode 2015–2024. Ditemukan bahwa secara kuantitatif global, model K-Means dengan parameter optimal $K=2$ menghasilkan rata-rata Silhouette Score yang sedikit lebih tinggi (0.502) dibandingkan model DBSCAN terbaik (0.468). Namun, keunggulan numerik ini terbukti semu karena evaluasi validitas internal mendeteksi adanya kesalahan pengelompokan (misclassification), yang ditandai dengan munculnya nilai siluet negatif pada wilayah dengan frekuensi bencana ekstrem, seperti Kabupaten Bandung dan Kota Bogor. Hal ini secara kuat menegaskan keterbatasan rigiditas K-Means dalam menangani data yang mengandung banyak outlier. Sebaliknya, algoritma DBSCAN terbukti lebih superior secara kualitatif dan representatif dalam memetakan risiko bencana yang kompleks. Meskipun memiliki skor rata-rata yang sedikit lebih rendah, DBSCAN menunjukkan kemampuan unggul dalam mempertahankan struktur alami data dengan cara mengisolasi enam wilayah ekstrem (22% data) sebagai noise, yang justru memberikan gambaran yang lebih akurat mengenai struktur data yang heterogen dan tidak seragam di Jawa Barat. Keunggulan adaptif ini semakin diperkuat oleh temuan analisis sensitivitas temporal, di mana DBSCAN mampu menangkap dinamika peningkatan ketidakpastian risiko bencana, yang ditunjukkan oleh lonjakan jumlah wilayah noise dari 5 pada periode 2015–2019 menjadi 10 wilayah pada periode 2020–2024. Peningkatan ini mengindikasikan bahwa pola bencana di Jawa Barat telah menjadi semakin tidak teratur dalam lima tahun terakhir. Oleh karena itu, penelitian ini merekomendasikan DBSCAN sebagai metode yang lebih robust dan andal untuk analisis spasial kebencanaan dan deteksi anomali risiko sejak dini.

REFERENCES

- [1] D. Ruslanjari, E. W. Safitri, F. A. Rahman, and C. Ramadhan, "ICT for public awareness culture on hydrometeorological disaster," *International Journal of Disaster Risk Reduction*, vol. 92, p. 103690, 2023, doi: <https://doi.org/10.1016/j.ijdr.2023.103690>.
- [2] Dikdik Ripaldi, "Mayoritas Daerah di Jabar Rawan Bencana, Dinkes Susun Rencana Antisipasi Krisis Kesehatan - Regional Liputan6.com." [Online]. Available: <https://www.liputan6.com/regional/read/5325144/mayoritas-daerah-di-jabar-rawan-bencana-dinkes-susun-rencana-antisipasi-krisis-kesehatan>. [Accessed: 28-Oct-2025].
- [3] H. J. Christanto, S. A. Sutresno, Y. A. Singgalen, and C. Dewi, "Analyzing Benefits of Online Train Ticket Reservation App Using Technology Acceptance Model," *Ingenierie des Systemes d'Information*, vol. 29, no. 1, pp. 107–115, Feb. 2024, doi: 10.18280/isi.290112.
- [4] C. Dewi and H. J. Christanto, "Automatic medical face mask recognition for COVID-19 Mitigation: Utilizing YOLO V5 object detection," *Revue d'Intelligence Artificielle*, vol. 37, no. 3, pp. 627–638, Jun. 2023, doi: 10.18280/ria.370312.
- [5] A. Baldah, V. Duarisah, and R. A. Maulana, "Clustering Daerah Rawan Bencana Alam Di Indonesia Berdasarkan Provinsi Dengan Metode K-Means," *Jurnal Ilmiah Informatika Global*, no. 2, pp. 31–36, Aug. 2023, [Online]. Available: <https://www.bps.go.id>. [Accessed: 3-Nov-2025].
- [6] M. Firman, A. Halik, and L. Septiana, "Analisa Data Untuk Prediksi Daerah Rawan Bencana Alam Di Jawa Barat Menggunakan Algoritma K-Means Clustering," *Journal of Information System, Applied, Management, Accounting and Research*, vol. 6, no. 4, pp. 856–870, 2022, doi: 10.52362/jisamar.v6i4.939.
- [7] J. Homepage, M. Herviany, S. Putri Delima, and T. Nurhidayah, "MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison of K-Means and K-Medoids Algorithms for Grouping Landslide Prone Areas in West Java Province Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Daerah Rawan Tanah Longsor di



- Provinsi Jawa Barat,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 1, pp. 34–40, Apr. 2021.
- [8] D. A. N. Sirodj, M. N. Aidi, B. Sartono, U. D. Syafitri, and B. Pranata, “Identification of Earthquake Prone Zones in Sumatra using Density Based Spatial Clustering of Applications with Noise,” *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 10, no. 2, pp. 916–928, Sep. 2025, doi: 10.18860/cauchy.v10i2.36120.
- [9] F. Rakhmawati and A. History, “Clustering flood prone areas in deli serdang reGENCY using density-based spatial clustering of applications with noise (dbscan) method,” *Desimal: Jurnal Matematika*, vol. 6, no. 2, pp. 261–272, Aug. 2023, doi: 10.24042/djm.
- [10] R. Adha, N. Nurhaliza, and U. Soleha, “Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia,” *Jurnal Sains, Teknologi dan Industri*, vol. 18, no. 2, pp. 206–211, 2021, [Online]. Available: <https://covid19.who.int>. [Accessed: 7-Nov-2025].
- [11] J. Li, A. Zheng, W. Guo, N. Bandyopadhyay, Y. Zhang, and Q. Wang, “Urban flood risk assessment based on DBSCAN and K-means clustering algorithm,” *Geomatics, Natural Hazards and Risk*, vol. 14, no. 1, pp. 1–28, May 2023, doi: 10.1080/19475705.2023.2250527.
- [12] R. A. Casonatto, T. D. P. G. Souza, and A. M. Mariano, “Quality and Risk Management in Data Mining: A CRISP-DM Perspective,” *Procedia Comput Sci*, vol. 242, pp. 161–168, Jan. 2024, doi: 10.1016/J.PROCS.2024.08.257.
- [13] M. F. A. Faqih and E. Mailoa, “Clustering zonasi daerah rawan bencana alam Provinsi Jawa Tengah menggunakan algoritma k-means dan library geopandas,” *IT-Explore: Jurnal Penerapan Teknologi Informasi dan Komunikasi*, vol. 4, no. 1, pp. 116–127, Feb. 2025, doi: 10.24246/itexplore.v4i1.2025.pp116-127.
- [14] “Normalisasi Data: Pengertian, Tujuan, dan Metodenya - Trivusi.” [Online]. Available: <https://www.trivusi.web.id/2022/09/normalisasi-data.html>. [Accessed: 8-Nov-2025].
- [15] A. Suarisman, A. Nazir, F. Syafrina, and L. Afriyanti, “Perbandingan Jarak Metrik pada Klasifikasi Jamur Beracun Menggunakan Algoritma K-Nearest Neighbor (K-NN),” *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 1, pp. 10–19, Nov. 2023, doi: 10.47065/josyc.v5i1.4511.
- [16] P. Palinggik Allorerung, A. Erna, M. Bagussahrir, and S. Alam, “Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit,” *Jurnal Informatika Sunan Kalijaga*, vol. 9, no. 3, pp. 178–191, Sep. 2024, doi: 10.14421/jiska.2024.9.3.178-191.
- [17] Y. Hasija, “Statistical methods in bioinformatics,” *All About Bioinformatics*, pp. 43–75, 2023, doi: 10.1016/B978-0-443-15250-4.00009-5.
- [18] I. J. Golub, R. M. Vakharia, and A. E. Razi, “Basic science statistics,” *Translational Orthopedics*, pp. 163–166, Jan. 2024, doi: 10.1016/B978-0-323-85663-8.00048-9.
- [19] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhajja, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf Sci (N Y)*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/J.INS.2022.11.139.
- [20] T. Amalina, D. Bima, A. Pramana, and B. N. Sari, “Metode K-Means Clustering Dalam Pengelompokan Penjualan Produk Frozen Food,” *Jurnal Ilmiah Wahana Pendidikan*, vol. 8, no. 15, pp. 574–583, Aug. 2022, doi: 10.5281/zenodo.7052276.
- [21] U. W. Latifah, S. Bahri, and M. Satriandhini, “Implementasi Algoritma K-Means Clustering untuk Strategi Promosi Kampus IBISA,” *JIKO (Jurnal Informatika dan Komputer)*, vol. 8, no. 2, p. 292, Sep. 2024, doi: 10.26798/jiko.v8i2.1307.
- [22] F. Handayani, “Aplikasi Data Mining Menggunakan Algoritma K-Means Clustering untuk Mengelompokkan Mahasiswa Berdasarkan Gaya Belajar,” *Jurnal Teknologi dan Informasi*, vol. 12, no. 1, pp. 46–63, Mar. 2022, doi: 10.34010/jati.v12i1.
- [23] J. Homepage *et al.*, “MALCOM: Indonesian Journal of Machine Learning and Computer Science Clustering Electricity Distribution Data Using Density-Based Spatial Clustering of Applications With Noise (DBSCAN) Algorithm Pengelompokan Data Pendistribusian Listrik Menggunakan Algoritma Density Based Spatial Clustering of Application With Noise (DBSCAN),” vol. 4, no. 3, pp. 1024–1033, 2024, doi: 10.57152/malcom.v4i2.1426.
- [24] I. N. Simbolon and P. D. Friskila, “ANALISIS DAN EVALUASI ALGORITMA DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE) PADA TUBERKULOSIS,” *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3S1, Oct. 2024, doi: 10.23960/jitet.v12i3s1.5206.
- [25] Fajar Supriadi, “Penerapan Metode Dbscan Dalam Penentuan Mahasiswa Yang Layak Memperoleh Bidikmisi,” *Management of Information System Journal*, vol. 3, no. 2, pp. 46–60, Mar. 2025, doi: 10.47065/mis.v2i3.1943.
- [26] P. Aselnino and A. W. Wijayanto, “Analisis Perbandingan Metode Hierarchical dan Non-Hierarchical dalam Pembentukan Cluster Provinsi di Indonesia Berdasarkan Indikator Women Empowerment,” *Indonesian Journal of Applied Statistics*, vol. 6, no. 1, p. 57, Jan. 2024, doi: 10.13057/ijas.v6i1.68876.
- [27] D. Hartama and S. Oktaviani, “OPTIMIZATION OF K-MEANS AND K-MEDOIDS CLUSTERING USING DBI SILHOUETTE ELBOW ON STUDENT DATA,” *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 11, no. 2, pp. 289–296, Mar. 2025, doi: 10.33330/jurteksi.v11i2.3531.
- [28] M. R. Irianto, A. Maududie, and F. N. Arifin, “Implementation of K-Means Clustering Method for Trend Analysis of Thesis Topics (Case Study: Faculty of Computer Science, University of Jember),” *BERKALA SAINSTEK*, vol. 10, no. 4, p. 210, Dec. 2022, doi: 10.19184/bst.v10i4.29524.