PENERAPAN ALGORITMA *DENSITY-BASED SPATIAL CLUSTERING OF APPLICATION WITH NOISE* UNTUK ANALISIS MULTIDIMENSI DATA DESTINASI WISATA PROVINSI JAWA TENGAH

Kholidatus Emilia[1], Kartika Maulida Hindrayani[2*], Alfan Rizaldy Pratama[3]
[1] Sains Data, Universitas Pembangunan Nasional Veteran Jawa Timur, Kota Surabaya, Jawa Timur, Indonesia

**Penulis korespondensi: kartika.maulida.ds@upnjatim.ac.id*

ABSTRAK

Pariwisata merupakan salah satu sektor strategis dalam pembangunan ekonomi Jawa Tengah, namun pemanfaatan data destinasi wisata yang semakin besar jumlahnya dan multidimensi belum sepenuhnya mendukung perencanaan yang merata. Ketimpangan kunjungan dan fasilitas antara destinasi populer dan destinasi yang kurang terekspos menunjukkan perlunya pemetaan klaster berbasis data untuk mendukung kebijakan yang lebih tepat. Penelitian ini bertujuan menganalisis pola pengelompokan 1.103 destinasi wisata di Jawa Tengah menggunakan pendekatan *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) pada pemodelan data multidimensi. Tahapan penelitian meliputi integrasi data, rekayasa fitur menjadi beberapa subfitur dan fitur gabungan, optimasi parameter *epsilon* dan jumlah titik minimum (MinPts), serta evaluasi klaster dengan beberapa metrik evaluasi. Hasil penelitian menunjukkan bahwa subfitur geografis menjadi model global terbaik dengan nilai evaluasi tinggi dan noise rendah. Pada pemodelan multidimensi, penggabungan fitur dengan penerapan *Principal Component Analysis* (PCA) menghasilkan lima klaster kualitas baik (Silhouette = 0.704; DBI = 0.866; CHI = 658.889), sedangkan fitur gabungan tanpa reduksi dimensi menunjukkan pemisahan klaster yang lebih lemah (Silhouette = 0.359; DBI = 2.124; CHI = 201.586). Temuan ini menegaskan bahwa algoritma DBSCAN, dengan parameter yang sistematis dan reduksi dimensi, efektif untuk mengungkap struktur kepadatan serta memberikan kerangka segmentasi destinasi yang objektif untuk mendukung kebijakan pemerataan pengembangan pariwisata di Provinsi Jawa Tengah.

Kata kunci: DBSCAN, Clustering, Data Multidimensi, Pariwisata, Jawa Tengah.

1 PENDAHULUAN

Pariwisata merupakan sektor penting bagi perekonomian daerah karena mampu mendorong aktivitas ekonomi lintas sektor dan menciptakan peluang positif bagi pendapatan masyarakat maupun pemerintah daerah (Rachman & Huda, 2025). Di antara berbagai provinsi di Indonesia, Jawa Tengah menempati posisi ketiga sebagai provinsi tujuan dengan jumlah perjalanan wisatawan nusantara tertinggi (Kemenparekraf RI, 2024). Dinamika pariwisata di Provinsi Jawa Tengah berkembang pesat seiring bertambahnya jumlah destinasi dan informasi pendukungnya. Hal ini dibuktikan dengan kajian Badan Pusat Statistik (BPS) Provinsi Jawa Tengah yang menunjukkan pertumbuhan PDRB sektor pariwisata 2013–2023 berfluktuasi, sempat terkontraksi tajam pada 2020 (-20,69%) akibat pandemi, lalu pulih hingga 7,54% pada 2023. Pada saat yang sama, kontribusi pariwisata antar kabupaten/kota menunjukkan kesenjangan misalnya Kota Semarang tercatat paling tinggi, sedangkan beberapa daerah berada pada level jauh lebih rendah dan jumlah daya tarik wisata, hotel, serta restoran juga menyebar secara tidak merata. Pola ini

berpotensi memperlebar ketimpangan pengembangan destinasi apabila tidak ditopang intervensi kebijakan yang lebih terarah dan berbasis data (BPS Provinsi Jawa Tengah, 2024).

Perkembangan digitalisasi telah memperkaya variasi data pariwisata secara signifikan, mulai dari data spasial seperti koordinat geografis dan aksesibilitas, data fasilitas, hingga data persepsi wisatawan berupa rating dan ulasan (Yolandari & Lastri Elisabet Butarbutar, 2025). Berbagai literatur menunjukkan bahwa pemanfaatan big data dan analitik memiliki potensi besar untuk mendukung pengambilan keputusan strategis di sektor pariwisata (Mariani et al., 2018). Sayangnya, penelitian di bidang ini masih cenderung tidak terorganisir, sehingga diperlukan pendekatan yang lebih terintegrasi agar hasil analisis dapat membentuk kebijakan praktis yang adil dan merata. Dalam perspektif keberlanjutan, big data analytics juga dinilai krusial untuk membantu pemilihan destinasi, meningkatkan pengalaman wisatawan, serta menyediakan dasar pengambilan keputusan bagi para pemangku kepentingan (Agrawal et al., 2022). Urgensi ini sejalan dengan Indeks Pembangunan Kepariwisataan Nasional (IPKN) 2024 di Indonesia, yang menempatkan ketersediaan dan kebaruan data sebagai salah satu pilar utama tata kelola. Hal ini menegaskan bahwa prioritas pengembangan pariwisata nasional harus didasarkan pada data yang lengkap dan terbaru (Kemenparekraf RI, 2024).

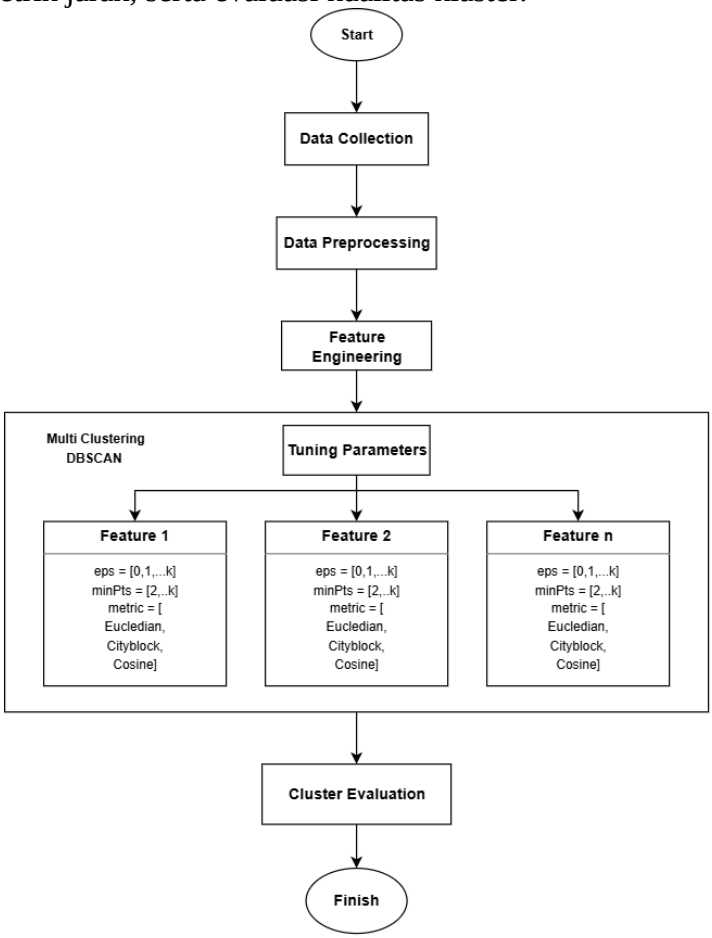
Untuk menjawab kebutuhan tersebut, analisis pengelompokan (*clustering*) adalah pendekatan yang paling logis yang bisa diterapkan, karena mampu mengidentifikasi segmen destinasi yang mirip berdasarkan pola tertentu, termasuk pola spasial (Maheswari et al., 2025). Penelitian terdahulu mengimplementasikan algoritma K-Means untuk mengelompokkan kabupaten/kota di Jawa Tengah berdasarkan potensi pariwisatanya (Maulidia & Astuti, 2022). Akan tetapi, metode berbasis centroid seperti K-Means memiliki keterbatasan jika diterapkan pada data destinasi wisata yang bersifat spasial dan heterogen, yaitu keharusan menentukan jumlah klaster di awal dan sensitivitasnya terhadap *outlier* yang dapat menggeser pusat klaster. Sebaliknya, algoritma *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) dinilai lebih sesuai untuk karakteristik data tersebut karena bekerja berdasarkan kepadatan data (*density-based*). Keunggulan DBSCAN terletak pada kemampuannya mendeteksi klaster dengan bentuk yang tidak beraturan serta memisahkan *noise* atau *outlier* secara otomatis. Efektivitas algoritma DBSCAN telah dibuktikan pada penelitian yang dilakukan oleh Fauzan dkk. (2022) untuk memetakan pola sebaran hotel di Bali yang menghasilkan klaster berkualitas dengan nilai *silhoutte* 0,709 dan penelitian Ramadhan dkk. (2025) untuk segmentasi UKM pariwisata di Rembang, di mana DBSCAN terbukti menghasilkan klaster yang lebih berkualitas dengan nilai evaluasi lebih tinggi dibandingkan K-Means (Fauzan et al., 2022; Ramadhan et al., 2025). Meskipun demikian, tantangan utama metode ini adalah penentuan parameter Epsilon dan MinPts. Kedua parameter ini sangat sensitif, terutama pada data dengan variasi kepadatan tinggi, sehingga pemilihannya harus dilakukan secara sistematis, misalnya menggunakan metode k-distance graph dan evaluasi validitas internal (Faizal et al., 2025).

Berdasarkan uraian permasalahan dan keterbatasan sebelumnya, penelitian ini akan menerapkan algoritma DBSCAN untuk memetakan klaster 1.103 destinasi wisata di Provinsi Jawa Tengah secara multidimensi, dengan asumsi bahwa destinasi membentuk kepadatan berbeda yang dipengaruhi kombinasi berbagai dimensi data. Implementasi dilakukan melalui rekayasa fitur menjadi beberapa subfitur dan satu fitur gabungan. Beberapa fitur akan dinormalisasi agar fitur berada pada skala yang sebanding dan mencegah dominasi fitur tertentu, sedangkan untuk fitur gabungan dilakukan reduksi dimensi dengan *Principal Component Analysis* (PCA). Kedua parameter DBSCAN akan ditentukan dan dioptimalkan secara sistematis berbasis dimensi data dan kurva *k-distance*, kemudian dievaluasi dengan *Silhouette Score*, *Davies–Bouldin Index,* dan

Calinski–Harabasz Index. Dengan pendekatan ini, penelitian diharapkan memberikan dasar segmentasi destinasi yang lebih objektif untuk mendukung perencanaan, pemerataan pengembangan, dan penentuan prioritas intervensi kebijakan pariwisata di Jawa Tengah.

2 METODE

Penelitian ini melalui beberapa tahap sistematis agar menghasilkan klasterisasi yang akurat dengan berbagai atribut yang relevan. Menggunakan pendekatan *data mining* dengan *unsupervised learning* untuk mengelompokkan destinasi wisata di Provinsi Jawa Tengah berdasarkan karakteristik multidimensi. Proses penelitian disusun secara bertahap dengan alur yang disajikan pada Gambar 1. agar menghasilkan klaster yang stabil dan interpretatif, meliputi: pengumpulan data, pra-pemprosesan, rekayasa fitur, multi-clustering menggunakan DBSCAN dengan penalaan parameter dan variasi metrik jarak, serta evaluasi kualitas klaster.



Gambar 1. Alur Penelitian

2.1 Data Collection

Data yang digunakan dalam penelitian ini dikumpulkan melalui proses *web scraping* Google Maps, karena platform tersebut menyediakan informasi *point of interest* (POI) yang relatif konsisten lintas wilayah, seperti: rating, jumlah ulasan, serta koordinat lintang-bujur. Jumlah record data yang berhasil diambil yaitu sekitar 1103 destinasi wisata. Setiap baris merepresentasikan satu destinasi wisata di Provinsi Jawa Tengah yang mencakup 35 kabupaten/kota, sehingga dataset dapat digunakan untuk analisis spasial dan pengelompokan destinasi secara komprehensif. Selain mengambil informasi dari POI, proses *web scraping* juga

mengambil data dengan dimensi yang berbeda dari sebelumnya, yaitu mengambil data dengan mencari fasilitas seperti penginapan dan hotel terdekat via *search nearby*. Dengan logika program akan mencari fasilitas penginapan terdekat, masih berada di sekitar destinasi wisata dengan radius maksimum 20 km. Hal yang sama juga dilakukan untuk mencari tempat makan terdekat dengan radius maksimal 10 km. Sehingga, jika digabungkan keseluruhan akan memuat beberapa atribut identitas dan atribut numerik yang relevan untuk pemodelan klaster dengan berbagai dimensi data. Rincian mengenai data yang digunakan dalam penelitian ini disajikan pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Keterangan
x_1	Garis lintang (*latitude*)
x_2	Garis bujur (*longitude*)
x_3	Rating destinasi wisata di Google Maps
x_4	Jumlah ulasan wisata di Google Maps
x_5	Jumlah penginapan di sekitar destinasi wisata dengan radius maksimal 20 km
x_6	Jumlah tempat makan di sekitar destinasi wisata dengan radius maksimal 10 km

2.2 Data Preprocessing

Tahap pra-pemprosesan dilakukan untuk memastikan kualitas dan kesiapan data sebelum digunakan pada algoritma berbasis jarak seperti DBSCAN (). Karena pembentukan klaster sangat bergantung pada kedekatan antartitik dan stabilitas parameter jarak (*eps*) yang digunakan. Langkah-langkah utama pra-pemrosesan meliputi:

2.2.1 Validasi struktur data

Data diperiksa untuk memastikan tidak ada nilai *null* atau data yang hilang, konsistensi tipe data, pembersihan duplikasi dari duplikasi data, serta kewajaran rentang nilai.

2.2.2 Transformasi untuk mengurangi skewness pada variabel

Variabel berbentuk hitungan seperti jumlah penginapan dan jumlah tempat makan cenderung memiliki sebaran yang tidak merata (*skewed*). Oleh karena itu digunakan transformasi logaritmik ringan dengan persamaan pada equation 1.

$$x' = \log(1 + x) \tag{1}$$

Transformasi ini membantu menurunkan dominasi nilai ekstrim tanpa menghilangkan makna ordinal seperti jumlah fasilitas.

2.2.3 Standarisasi fitur numerik (feature scaling)

Karena DBSCAN sangat dipengaruhi skala data (melalui perhitungan jarak), fitur numerik distandarisasi menggunakan StandardScaler (*z-score*). *z-score* atau skor baku merupakan suatu nilai yang mengukur jarak suatu data dengan nilai asli (x) dari nilai rata-rata (μ), yang dinyatakan dalam satuan simpangan baku (σ) (Aziz et al., 2021). *z-score* dapat dicari dengan persamaan 2.

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Praktik standarisasi dengan StandardScaler juga umum digunakan sebagai bagian dari *feature scaling* sebelum pemodelan.

2.3 Feature Engineering

Pada penelitian ini, *feature engineering* digunakan untuk merapikan dan menyetarakan berbagai atribut pada dataset wisata Jawa Tengah agar siap diproses algoritma DBSCAN. DBSCAN bekerja dengan membentuk klaster dari kepadatan titik yang dihitung lewat jarak, jika ada fitur yang skalanya jauh lebih besar, jarak akan "ditarik" oleh fitur itu dan hasil klaster bisa bias (Wang et al., 2019). Karena itu, semua fitur numerik yang dipakai distandarisasi ke skala yang

sebanding, sehingga tiap fitur punya peluang kontribusi yang lebih merata dalam perhitungan jarak (Faizal et al., 2025).

Subfitur pertama adalah '***geo_features'***, yaitu koordinat geografis latitude dan longitude yang akan dilakukan standarisasi terlebih dahulu supaya tidak ada koordinat lebih "mendominasi" hanya karena beda rentang angka. Proses standarisasi menggunakan StandarScaler dengan persamaan 3.

$$geo\_features_i = \frac{x_i - \mu_i}{\sigma_i}, for\ i\ \in \{latitude, longitude\} \quad (3)$$

Dimana x_i adalah nilai asli dari fitur geografis, μ_i adalah rata-rata, dan σ_i adalah simpangan baku dari fitur geografis.

Subfitur kedua, '***facilities_features***' dibentuk dari jumlah fasilitas sekitar destinasi wisata seperti penginapan dan tempat makan dengan batasan maksimum radius 20 km untuk penginapan dan 10 km untuk tempat makan, batasan ini digunakan agar logika pencarian fasilitas tidak keluar dari batas atas yang sudah ditetapkan dan berakhir menggunakan fasilitas dengan jarak yang jauh dari destinasi wisata. Nilai pada fitur ini akan ditransformasi logaritmik ringan dengan persamaan pada equation 1. kemudian distandarisasi agar tidak ada perbedaan ekstrim antar data.

Subfitur ketiga, '***popularity_features***' yang merepresentasikan skor 'rating' sebagai indikator kepuasan pelanggan ditransformasikan menjadi variabel kategorikal bertingkat melalui proses *binning*. Skema binning disesuaikan dengan karakteristik dataset *Google Rating* yang Sebagian besar berada pada rentang tinggi sekitar 4.0–4.8, sehingga rating diklasifikasikan menjadi tiga kategori tingkat popularitas sebagaimana dirumuskan pada persamaan (4). Selanjutnya, kategori hasil binning dikonversi menggunakan *one-hot encoding* menjadi fitur biner dengan Tingkat popularitas 'Buruk', 'Baik', dan 'Baik_Sekali' agar dapat direpresentasikan secara eksplisit dan konsisten dalam proses klasterisasi.

$$pop_i = \begin{cases} Buruk, if\ rating_i \leq 3.9 \\ Baik, if\ 4.0 < rating_i \leq 4.4 \\ Baik\ Sekali, if\ rating_i > 4.4 \end{cases} \quad (4)$$

Subfitur keempat, '*segment_features*' merupakan segmentasi destinasi wisata dari kombinasi nilai *rating* dan *jumlah_ulasan*. Logika dasar yang digunakan yaitu: rating menggambarkan kualitas persepsi pengunjung, sedangkan jumlah ulasan dipakai sebagai indikasi intensitas atensi/keramaian. Keduanya kemudian dibagi menjadi dua tingkat (*High* dan *Low*) menggunakan ambang statistik (median setelah transformasi dengan persamaan 1. untuk jumlah_ulasan agar tidak terlalu bias oleh nilai ekstrim). Dari kombinasi dua tingkat ini terbentuk empat segmen: rating_high_only (rating tinggi tetapi ulasan rendah, destinasi cenderung "disukai namun belum ramai"), reviews_high_only (ulasan tinggi tetapi rating rendah, "ramai tetapi kualitas persepsinya biasa saja"), both_high (rating dan ulasan sama-sama tinggi, "ramai dan disukai"), serta both_low (keduanya rendah, "tidak ramai dan biasa saja").; ketika keduanya dipadukan, kita bisa membedakan destinasi yang "disukai tapi belum ramai", "ramai tapi biasa saja", "tidak ramai dan biasa saja", sampai "ramai dan sangat disukai" perumusan segmentasi pada persamaan 5. Konsep segmentasi lalu di-*one-hot encoding* agar bisa dipakai DBSCAN sebagai fitur numerik. Sebelum melakukan segmentasi, akan didefinisikan terlebih dahulu indikator "tinggi" untuk rating dan ulasan:

$$R_i = \begin{cases} 1, & r_i > \theta_r \\ 0, & r_i \leq \theta_r \end{cases}, \qquad U_i = \begin{cases} 1, & g(u_i) > \theta_r \\ 0, & g(u_i) \leq \theta_r \end{cases} \quad (5a)$$

Penjelasan:

- r_i = nilai rating
- u_i = jumlah ulasan
- θ_r = ambang rating
- θ_u = ambang ulasan

$$seg_i = \begin{cases} \text{rating\_high\_only}, & R_i = 1 \ \wedge U_i = 0 \\ \text{reviews\_high\_only}, & R_i = 0 \ \wedge U_i = 1 \\ \text{both\_high}, & R_i = 1 \ \wedge U_i = 1 \\ \text{both\_low}, & R_i = 0 \ \wedge U_i = 0 \end{cases} \quad (5)$$

Interpretasi singkatnya:

- $R_i = 1, U_i = 0$: "disukai tapi belum ramai"
- $R_i = 0, U_i = 1$: "ramai tapi biasa saja"
- $R_i = 1, U_i = 1$: "ramai dan disukai"
- $R_i = 0, U_i = 0$: "tidak ramai dan biasa saja"

Setelah tiap subfitur fitur dibuat, semuanya digabung menjadi '*feature_combined_raw'*, yaitu satu matriks fitur yang mewakili profil destinasi dari sisi lokasi, fasilitas sekitar, popularitas, dan segmen. Penggabungan ini berguna untuk melihat pola klaster "secara utuh", tetapi juga punya risiko ketika tipe informasi yang dicampur makin beragam, struktur data makin kompleks dan kualitas klaster bisa turun bila tidak ditangani dengan baik (Faizal et al., 2025). Karena itu, dibuat subfitur terakhir yaitu '*feature_combined_pca'* yang bekerja dengan menstandarisasi subfitur *feature_combined_raw*, lalu direduksi dimensinya memakai *Principal Component Analysis* (PCA) dengan target mempertahankan ≥95% variasi informasi. Sederhananya PCA "meringkas" banyak kolom yang saling berkorelasi menjadi beberapa komponen utama yang lebih padat, sehingga DBSCAN bekerja di ruang fitur yang lebih ringkas dan sering kali lebih bersih dari redundansi/noise. Pendekatan menggabungkan DBSCAN dengan PCA juga pernah diuji pada penelitian Mustakim dkk. (2021) dengan hasil peningkatan kualitas klaster yaitu nilai silhouette lebih tinggi dibanding DBSCAN tanpa reduksi dimensi, karena struktur data menjadi lebih mudah dipisahkan.

2.4 Multi Clustering DBSCAN

2.4.1 Parameter Tuning

Penentuan parameter DBSCAN pada penelitian ini dilakukan secara sistematis agar hasil klaster tidak bergantung pada *trial-and-error* semata. parameter utama yang dioptimasi adalah ε (eps) sebagai batas jarak maksimum, serta MinPts sebagai jumlah minimum titik untuk membentuk klaster. Nilai ε ditaksir melalui kurva k-distance, yaitu menghitung jarak tetangga ke-*k* (*K-nearest neighbor distance*) untuk setiap titik, mengurutkannya, kemudian memilih titik siku (elbow) sebagai kandidat ε yang paling representatif. Deteksi titik siku dapat dirumuskan dengan persamaan 6, menghitung jarak tegak lurus suatu titik pada kurva terhadap garis yang menghubungkan titik awal–akhir kurva, sehingga perubahan gradien yang paling tajam menjadi indikator ε yang optimal.

$$d_{line}(x) = \frac{|(x - p_1)\bar{v}|}{|\bar{v}|} \quad (6)$$

Selanjutnya, MinPts ditentukan menggunakan heuristik berbasis dimensi fitur dengan persamaan 7, dengan d menyatakan banyaknya dimensi pada subfitur fitur yang di klasterisasi.

$$minPts = \max(2 \times d, 4) \quad (7)$$

Dalam implementasinya, kandidat hasil perhitungan ε dan MinPts kemudian dieksplorasi melalui grid search pada setiap subfitur.

2.4.2 Distance metrics

Karena DBSCAN merupakan algoritma berbasis jarak, penelitian ini mengevaluasi hasil klaster dengan beberapa metrik jarak untuk menilai ketahanan (*robustness*) klaster terhadap kedekatan. Pengujian beberapa metrik jarak dengan tujuan untuk mencari metrik yang sesuai dengan tipe fitur. Penelitian ini menggunakan tiga metrik jarak yang meliputi:

Euclidean distance, untuk mengukur jarak garis lurus antar dua titik pada ruang berdimensi-n, persamaan 7 adalah rumus untuk mengukur kemiripan data dengan metrik *euclidean*.

$$d_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{7}$$

Cityblock/Manhattan distance, metrik ini bekerja dengan menjumlahkan selisih absolut setiap dimensi, sering lebih stabil pada data berdimensi lebih tinggi atau ketika ingin mengurangi pengaruh deviasi ekstrim pada satu dimensi. Persamaan 8 adalah rumus untuk mengukur kemiripan data dengan metrik *cityblock/manhattan*.

$$d_{Cityblock}(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{8}$$

Cosine distance, metrik ini bekerja dengan mengukur *dissimilarity* berdasarkan sudut antar vektor, relevan ketika pola arah/komposisi fitur (misalnya fitur *dummy*/one-hot) lebih penting daripada besaran absolutnya. Persamaan 9 adalah rumus untuk mengukur kemiripan data dengan metrik *cosine*.

$$d_{Cosine}(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2}\sqrt{\sum_{i=1}^{n} {y_i}^2}} \tag{9}$$

2.5 Cluster Evaluation

Karena penelitian ini bersifat unsupervised dan tidak memiliki label kelas (*ground truth*), kualitas klaster dievaluasi menggunakan metrik validitas internal. Metrik ini menilai hasil klaster berdasarkan dua aspek utama, yaitu kekompakan (*cohesion*) di dalam klaster dan keterpisahan (*separation*) antarklaster. Berikut metrik internal yang digunakan penelitian ini untuk mengevaluasi klaster yang terbentuk:

2.5.1 Silhouette Score

Silhouette mengukur seberapa baik sebuah titik ditempatkan pada klasternya dibandingkan dengan klaster terdekat. Untuk setiap titik i, Silhouette menggunakan jarak rata-rata titik i ke semua titik pada klasternya $a(i)$dan jarak rata-rata minimum ke klaster lain terdekat $b(i)$, lalu dihitung dengan persamaan 10.

$$Silhouette\ Score = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{10}$$

Nilai Silhouette berada pada rentang $[-1,1]$. Semakin mendekati 1 berarti klaster makin kompak dan terpisah jelas, sedangkan nilai mendekati 0 mengindikasikan tumpang tindih antar klaster. Secara implementatif, definisi a dan b serta bentuk rumus ini juga konsisten dengan dokumentasi metrik *clustering* yang umum dipakai pada eksperimen berbasis Python. Dalam pemilihan model, Silhouette sering dipakai sebagai indikator pemisahan klaster yang intuitif dan mudah dibandingkan antar-skenario (Rahmawati et al., 2024).

2.5.2 Davies-Bouldin Index (DBI)

DBI menilai kualitas klaster dengan mempertimbangkan dua aspek utama, yaitu kepadatan/dispersi di dalam klaster dan jarak pemisah antar klaster. Persamaan 11 adalah rumus lengkap perhitungan DBI

$$\text{DBI} = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \quad 11$$

Nilai DBI yang lebih kecil menunjukkan kualitas klasterisasi yang lebih baik karena mengindikasikan klaster yang lebih kompak dan saling terpisah secara jelas, sehingga potensi tumpang tindih antar klaster menjadi rendah. Sebaliknya, DBI yang lebih besar mengindikasikan bahwa beberapa klaster cenderung berdekatan atau kurang terpisah, yang menandakan proses klasterisasi belum optimal (Mauludin et al., 2025).

2.5.3 Calinski-Harabasz Index (CHI)

CHI (sering disebut juga *Variance Ratio Criterion*) mengevaluasi rasio dispersi antar-klaster terhadap dispersi dalam klaster, dengan mempertimbangkan ukuran sampel dan jumlah klaster. Persamaan 12 adalah rumus evaluasi CHI

$$\text{CHI} = \frac{B_k}{W_k} \times \frac{n-k}{k-1} \quad 12$$

dengan B adalah dispersi antar-klaster dan Wdispersi dalam klaster. Nilai CHI yang lebih besar mengindikasikan pemisahan klaster yang lebih tegas dan struktur yang lebih terdefinisi (Putri & Gusmanely, 2022)

3 HASIL DAN PEMBAHASAN

Setelah seluruh tahapan metode dilakukan, bagian ini akan memaparkan hasil klasterisasi destinasi wisata Jawa Tengah untuk setiap subset fitur yang diuji. Ringkasan performa terbaik dan konfigurasi optimal disajikan pada Tabel 2, lalu pembahasan dilanjutkan dengan interpretasi profil klaster pada tabel berikutnya untuk menjelaskan perbedaan karakter destinasi antar-klaster.

3.1 Ringkasan kinerja terbaik tiap subfitur fitur

Tabel 2. Hasil kinerja terbaik dari setiap subfitur

Subfiturs	Optimal Parameter	n_cluster	n_noise	Sil_score	DBI	CHI
geo_features	*eps* = 0.025 *minPts* = 50 *metric* = cosine	3	30	0.767104	0.730477	1134.3956
facilities_features	*eps* = 0.0098 *minPts* = 20 *metric* = cosine	3	42	0.754121	0.875202	501.80493
popularity_features	*eps* = 0.5 *minPts* = 2 *metric* = cosine	3	0	1	0	inf
segment_features	*eps* = 0.707 *minPts* = 2 *metric* = euclidean	4	0	1	0	inf

feature_combined_pca	*eps* = 0.073 *minPts* = 16 *metric* = cosine	5	164	0.703671	0.866213	658.88913
feature_combined_raw	*eps* = 2.654 *minPts* = 22 *metric* = cityblock	2	64	0.359275	2.124075	201.58572

Berdasarkan Tabel 2, subfitur *geo_features* menghasilkan performa paling stabil dan paling kuat untuk pemetaan spasial: terbentuk 3 klaster dengan 30 noise, serta memperoleh Silhouette Score 0,767, DBI 0,730, dan CHI 1134,396. Kombinasi ini menunjukkan struktur klaster yang relatif kompak dan terpisah dengan baik, sementara proporsi noise yang rendah mengindikasikan mayoritas destinasi berada dalam kelompok-kelompok kepadatan spasial yang jelas. Subfitur *facilities_features* juga membentuk struktur klaster yang cukup baik (Silhouette 0,754) dengan 3 klaster dan 42 noise. Walaupun nilai Silhouette mendekati model spasial, akan tetapi nilai CHI lebih rendah (501,805) yang mengisyaratkan pemisahan antar-klaster dalam ruang "kepadatan fasilitas" tidak sebagus pemisahan spasial murni. Dengan kata lain, fasilitas di sekitar destinasi membentuk pola pengelompokan, tetapi variasinya lebih gradual dan saling tumpang tindih dibanding pola kedekatan geografis. Pada subfitur *popularity_features* dan *segment_features*, metrik evaluasi menunjukkan nilai sempurna (Silhouette ≈ 1, DBI ≈ 0, CHI ≈ ∞). Namun hasil ini perlu diterjemahkan dengan hati-hati, karena fitur yang digunakan berbasis kategori/*one-hot* sehingga kombinasi nilainya terbatas dan banyak titik menjadi identik. Kondisi tersebut dapat membuat klaster tampak terpisah total secara matematis, tetapi tidak selalu menggambarkan struktur kepadatan yang "alami" seperti pada data spasial atau fasilitas. Secara substantif, dua subfitur ini lebih tepat dipandang sebagai segmentasi berbasis kategori popularitas (rating/ulasan) daripada pemetaan kepadatan destinasi lintas dimensi.

Perbandingan penting lain terlihat pada fitur gabungan: *feature_combined_raw* menghasilkan kualitas klaster paling rendah (Silhouette 0,359, DBI 2,124) meskipun hanya membentuk 2 klaster. Ini mengindikasikan penggabungan fitur heterogen secara langsung cenderung menghasilkan ruang fitur yang "noise" dan sulit dipisahkan menggunakan DBSCAN. Sebaliknya, ketika fitur gabungan direduksi dengan PCA (*feature_combined_pca*), kualitas klaster meningkat cukup signifikan (Silhouette 0,704, DBI 0,866, CHI 658,889), meskipun konsekuensinya adalah noise meningkat (164). Temuan ini menguatkan bahwa reduksi dimensi membantu DBSCAN bekerja pada ruang fitur gabungan yang kompleks, namun sebagian destinasi tetap memiliki profil yang terlalu unik/tersebar sehingga tidak masuk ke klaster padat.

3.2 Profil klaster pada subfitur

3.2.1 Profil klaster pada *geo_features*

Tabel 3. Ringkasan Profil Klaster Subfitur *geo_features*

Cluster	Jumlah_ destinasi	Mean atribut
0	440	Rating_mean ≈ 4.2 Ulasan_mean ≈ 1314 JmlPenginapan_mean ≈ 24 JmlTempatMakan_mean ≈ 18

1	332	Rating_mean ≈ 4.1 Ulasan_mean ≈ 1898 JmlPenginapan_mean ≈ 23 JmlTempatMakan_mean ≈ 18
2	302	Rating_mean ≈ 4.1 Ulasan_mean ≈ 2773 JmlPenginapan_mean ≈ 26 JmlTempatMakan_mean ≈ 19

Tabel 3 merangkum profil klaster dari model spasial (geo_features). Tiga klaster terbentuk dengan ukuran yang relatif seimbang (Klaster 0: 440 destinasi; Klaster 1: 332; Klaster 2: 302). Walaupun klaster hanya dibangun dari koordinat, perbedaan rata-rata atribut non-spasial memberi petunjuk bahwa aglomerasi spasial berkaitan dengan tingkat keramaian destinasi dan kepadatan amenitas pendukung.

Klaster 0 memiliki rating rata-rata 4,2 dengan ulasan rata-rata 1314 serta kepadatan fasilitas yang hampir merata (penginapan 24; tempat makan 18). Klaster 1 menunjukkan pola yang mirip dari sisi rating (4,1), tetapi ulasan lebih tinggi (1898). Sementara itu, Klaster 2 memiliki ulasan rata-rata tertinggi (2773) dan amenitas sedikit lebih tinggi (penginapan 26; tempat makan 19). Perbedaan rating antar klaster relatif kecil (4,1–4,2), sehingga dalam konteks ini jumlah ulasan tampak lebih diskriminatif untuk membedakan "tingkat keramaian/popularitas" dibanding rating yang cenderung tinggi merata. Implikasinya, pemetaan spasial dengan DBSCAN tidak hanya menunjukkan pengelompokan lokasi destinasi yang berdekatan, tetapi juga membantu mengidentifikasi kawasan yang secara empiris lebih ramai dan lebih siap secara amenitas dibanding kawasan lain.

Keberadaan noise pada subfitur spasial juga penting, noise dapat dibaca sebagai destinasi yang secara lokasi relatif terpisah dari aglomerasi destinasi lain, misalnya tersebar di wilayah pinggiran/akses lebih sulit. Kelompok ini perlu dianalisis lebih lanjut karena intervensinya biasanya berbeda, bukan hanya sekadar perkuat promosi, tetapi bisa berupa penguatan akses, integrasi rute, atau dukungan fasilitas dasar agar destinasi tidak berdiri sendiri di luar ekosistem wisata utama.

3.2.2 Profil klaster pada *facilities_features*

Tabel 4. Ringkasan Profil Klaster Subfitur *facilities_features*

Cluster	Jumlah_ destinasi	Mean atribut
0	425	Rating_mean ≈ 4.2 Ulasan_mean ≈ 1798 JmlPenginapan_mean ≈ 33 JmlTempatMakan_mean ≈ 21
1	534	Rating_mean ≈ 4.2 Ulasan_mean ≈ 2328 JmlPenginapan_mean ≈ 18 JmlTempatMakan_mean ≈ 20

2	103	Rating_mean ≈ 4.0 Ulasan_mean ≈ 913 JmlPenginapan_mean ≈ 20 JmlTempatMakan_mean ≈ 7

Tabel 4 menampilkan klaster yang dibentuk dari kepadatan fasilitas sekitar destinasi (jumlah penginapan radius 20 km dan tempat makan radius 10 km). Klaster 0 (425 destinasi) dengan kepadatan penginapan tertinggi (33) dan tempat makan yang juga tinggi (21), dan dengan rating 4,2 dan ulasan 1798. Ini dapat diartikan sebagai kelompok destinasi yang berada pada lingkungan dengan dukungan amenitas kuat kondisi yang umumnya memudahkan wisatawan untuk tinggal lebih lama atau melakukan kunjungan berulang.

Klaster 1 (534 destinasi) menarik karena memiliki ulasan rata-rata paling tinggi (2328) dengan rating tetap tinggi (4,2), tetapi jumlah penginapan rata-rata lebih rendah (18) sementara tempat makan tetap tinggi (20). Pola ini mengindikasikan bahwa popularitas (ulasan tinggi) tidak selalu identik dengan kepadatan penginapan di sekitar destinasi; bisa terjadi pada destinasi yang kuat sebagai tujuan kunjungan harian (*day trip*) atau berada dekat pusat aktivitas masyarakat sehingga akomodasi tidak terkonsentrasi dalam radius yang sama. Dengan kata lain, klaster fasilitas membantu membedakan destinasi yang "ramai karena sistem akomodasinya" atau "ramai karena daya tariknya", dengan konsekuensi kebijakannya berbeda.

Klaster 2 (103 destinasi) menjadi kelompok yang paling perlu perhatian karena menunjukkan rating lebih rendah (4,0), ulasan paling rendah (913), dan terutama ketersediaan tempat makan sangat rendah (7). Ini dapat ditafsirkan sebagai destinasi dengan dukungan amenitas yang belum memadai sehingga pengalaman wisatawan berpotensi kurang optimal. Dalam konteks pemerataan pembangunan pariwisata, klaster ini bisa menjadi kandidat prioritas untuk intervensi bertahap—misalnya peningkatan layanan dasar, penguatan UMKM kuliner lokal, atau konektivitas ke pusat layanan terdekat.

3.2.3 Profil klaster pada *feature_combined_pca*

Tabel 5. Ringkasan Profil Klaster Subfitur *feature_combined_pca*

Cluster	Jumlah_ destinasi	Mean atribut
0	202	Rating_mean ≈ 4.5 Ulasan_mean ≈ 5652 JmlPenginapan_mean ≈ 26 JmlTempatMakan_mean ≈ 21
1	282	Rating_mean ≈ 4.3 Ulasan_mean ≈ 2587 JmlPenginapan_mean ≈ 25 JmlTempatMakan_mean ≈ 21
2	190	Rating_mean ≈ 4.3 Ulasan_mean ≈ 182 JmlPenginapan_mean ≈ 25 JmlTempatMakan_mean ≈ 20

3	192	Rating_mean ≈ 4.7 Ulasan_mean ≈ 131 JmlPenginapan_mean ≈25 JmlTempatMakan_mean ≈ 20
4	72	Rating_mean ≈ 1.2 Ulasan_mean ≈ 41 JmlPenginapan_mean ≈ 26 JmlTempatMakan_mean ≈ 18

Hasil pada Tabel 5 menunjukkan bahwa penggabungan dimensi (lokasi, fasilitas, dan popularitas/segmentasi) yang kemudian direduksi dimensi dengan PCA menghasilkan 5 klaster yang secara substantif lebih kaya untuk segmentasi destinasi. Klaster 0 (202 destinasi) merepresentasikan kelompok paling menonjol secara popularitas dengan rating 4,5 dan ulasan rata-rata sangat tinggi (5652), disertai fasilitas yang cukup stabil (penginapan 26; tempat makan 21). Kelompok ini dapat dibaca sebagai "destinasi unggulan" yang perlu pendekatan pengelolaan berbasis kapasitas (daya dukung), kualitas layanan, dan keberlanjutan, karena lonjakan permintaan wisata biasanya paling terkonsentrasi di segmen ini.

Klaster 1 (282 destinasi) menunjukkan profil "populer menengah" (rating 4,3; ulasan 2587), sedangkan Klaster 2 (190) dan Klaster 3 (192) sama-sama memiliki rating tinggi (4,3–4,7) tetapi ulasan rendah (182 dan 131). Dua klaster terakhir ini penting untuk strategi pemerataan destinasi dengan rating tinggi namun ulasan rendah berpotensi merupakan *hidden gems* atau destinasi yang kualitasnya baik tetapi eksposurnya belum kuat. Intervensinya cenderung efektif melalui penguatan promosi, integrasi paket wisata, serta peningkatan akses informasi digital, misalnya pengelolaan profil destinasi dan rute kunjungan

Klaster 4 (72 destinasi) menunjukkan rating rata-rata 1,2 dan ulasan 41. Nilai rating serendah ini tidak lazim pada data Google Maps untuk destinasi wisata, sehingga secara ilmiah temuan ini layak diperlakukan sebagai klaster anomali: bisa merepresentasikan destinasi yang memang bermasalah (pengalaman buruk, informasi tidak akurat, atau perubahan status operasional), namun bisa juga terkait isu kualitas data (misalnya nilai tidak terbaca/terisi dengan benar pada proses ekstraksi). Karena penelitian ini berorientasi pada rekomendasi kebijakan, klaster anomali seperti ini sebaiknya ditindaklanjuti dengan verifikasi sampel secara manual sebelum ditarik sebagai kesimpulan substantif.

Jumlah noise pada feature_combined_pca cukup tinggi. Dalam pembacaan algoritma DBSCAN, noise tidak selalu "data buruk", melainkan indikator bahwa sebagian destinasi memiliki profil multidimensi yang tidak membentuk kepadatan cukup kuat pada parameter yang dipilih. Misalnya destinasi unik dengan kombinasi fasilitas popularitas yang tidak banyak memiliki kesamaan di dataset. Secara praktis, noise dapat dimanfaatkan sebagai daftar prioritas analisis individual (*case-by-case*) karena pendekatan klaster tidak cukup mewakili karakter mereka.

3.2.4 Profil klaster pada *feature_combined_raw*

Tabel 6. Ringkasan Profil Klaster Subfitur *feature_combined_raw*

Cluster	Jumlah_ destinasi	Mean atribut
0	619	Rating_mean ≈ 4.0 Ulasan_mean ≈ 1417 JmlPenginapan_mean ≈ 25 JmlTempatMakan_mean ≈ 19
1	421	Rating_mean ≈ 4.6 Ulasan_mean ≈ 2922 JmlPenginapan_mean ≈ 25 JmlTempatMakan_mean ≈ 20

Tabel 6 menunjukkan bahwa pada fitur gabungan tanpa reduksi dimensi (raw), DBSCAN hanya membentuk 2 klaster utama dengan profil yang relatif "terlalu umum": Klaster 0 (619 destinasi) memiliki rating rata-rata 4,0 dan ulasan 1417, sedangkan Klaster 1 (421 destinasi) memiliki rating lebih tinggi (4,6) dengan ulasan 2922. Perbedaan jumlah penginapan dan tempat makan antar-klaster relatif kecil (sekitar 25–20), sehingga pemisahan klaster terutama didorong oleh popularitas, bukan oleh variasi fasilitas.

Temuan ini konsisten dengan kinerja evaluasi pada Tabel 2 yang menunjukkan Silhouette rendah (0,359) dan DBI tinggi (2,124): gabungan fitur heterogen tanpa reduksi dimensi cenderung membuat jarak antar titik kurang informatif, sehingga klaster yang terbentuk menjadi kurang tajam dan kurang kuat untuk segmentasi kebijakan. Dengan demikian, hasil ini memperkuat argumen bahwa untuk analisis multidimensi, pendekatan ***feature_combined_pca*** lebih layak dijadikan rujukan dibanding ***feature_combined_raw***, karena memberikan struktur klaster yang lebih granular dan interpretatif.

4 KESIMPULAN

Penelitian ini membuktikan bahwa penerapan DBSCAN pada data destinasi wisata Jawa Tengah dapat menghasilkan segmentasi yang informatif, terutama ketika klasterisasi difokuskan pada dimensi yang sesuai dengan prinsip kepadatan. Pada subfitur spasial (*geo_features*), DBSCAN menghasilkan tiga klaster dengan kualitas pemisahan paling baik (Silhouette 0.767; DBI 0.730; CHI 1134.396) dan noise yang rendah, sehingga efektif untuk memetakan dengan konsentrasi destinasi berdasarkan kedekatan geografis. Pada subfitur fasilitas (*facilities_features*), model juga membentuk tiga klaster dengan performa tinggi (Silhouette 0.754) yang mengindikasikan adanya perbedaan kepadatan fasilitas pendukung di sekitar destinasi wisata, pembentukan klaster pada dimensi ini dapat digunakan sebagai dasar prioritas pemerataan dukungan fasilitas antar destinasi.

Dalam pendekatan multidimensi, fitur gabungan tanpa reduksi dimensi menunjukkan kualitas klaster yang kurang tajam (Silhouette 0.359; DBI 2.124) dan hanya membentuk dua klaster, sementara penggunaan PCA pada fitur gabungan meningkatkan keterbacaan struktur klaster menjadi lima klaster (Silhouette 0.704; DBI 0,866) meskipun menghasilkan noise yang lebih besar, yang dapat diartikan sebagai indikasi heterogenitas profil destinasi atau ada keterbatasan kualitas data hasil scraping. Dengan demikian, kontribusi utama penelitian ini adalah memberikan kerangka segmentasi destinasi yang tidak hanya spasial, tetapi juga dapat diperluas secara multidimensi dan lebih stabil melalui penguatan representasi fitur.

Sebagai saran, penelitian selanjutnya dapat memperkaya fitur dengan data kunjungan resmi, teks ulasan, maupun indikator aksesibilitas yang lebih rinci, menerapkan metode density-based yang lebih adaptif terhadap variasi kepadatan sebagai pembanding, dan melakukan validasi eksternal pada sampel destinasi agar interpretasi klaster semakin kuat untuk kebutuhan kebijakan.

DAFTAR PUSTAKA

Agrawal, R., Wankhede, V. A., Kumar, A., Luthra, S., & Huisingh, D. (2022). Big data analytics and sustainable tourism: A comprehensive review and network based analysis for potential future research. *International Journal of Information Management Data Insights*, *2*(2). https://doi.org/10.1016/j.jjimei.2022.100122

Aziz, A. R., Warsito, B., & Prahutama, A. (2021). Pengaruh Transformasi Data Pada Metode Learning Vector Quantization Terhadap Akurasi Klasifikasi Diagnosis Penyakit Jantung. *Jurnal Gaussian*, *10*(1), 21–30. https://doi.org/10.14710/j.gauss.v10i1.30933

BPS Provinsi Jawa Tengah. (2024). *Kajian Ketimpangan Pembangunan Pariwisata Provinsi Jawa Tengah 2013-2023* (Vol. 3).

Faizal, E., Hartati, S., & Musdholifah, A. (2025). Multi-Cluster DBSCAN for Analysing Tourism Data. *International Journal of Intelligent Engineering and Systems*, *18*(1), 660–674. https://doi.org/10.22266/ijies2025.0229.47

Fauzan, A., Novianti, A., Ramadhani, R. R. M. A., & Adhiwibawa, M. A. S. (2022). Analysis of Hotels Spatial Clustering in Bali: Density-Based Spatial Clustering of Application Noise (DBSCAN) Algorithm Approach. *EKSAKTA: Journal of Sciences and Data Analysis*, 25–38. https://doi.org/10.20885/eksakta.vol3.iss1.art4

Kemenparekraf RI. (2024). *Indeks Pembangunan Kepariwisataan Nasional VOL.2, 2024*. https://www.kemenpar.go.id/berita/indeks-pembangunan-kepariwisataan-nasional-ipkn-vol-2-tahun-2024

Maheswari, A. D., Setiyorini, H. P. D., & Khaerani, R. (2025). Segmentasi Wisatawan Di Media Sosial Berbasis Klaster: Bagaimana Wisatawan Berinteraksi Dengan Konten Perjalanan Wisata? *Kepariwisataan: Jurnal Ilmiah*, *19*(2), 152. https://doi.org/10.47256/kji.v19i2.747

Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*, *30*(12), 3514–3554. https://doi.org/10.1108/IJCHM-07-2017-0461

Maulidia, S., & Astuti, E. T. (2022). Pengelompokan Kabupaten/Kota di Provinsi Jawa Tengah Berdasarkan Potensi Pariwisata Tahun 2020. *Seminar Nasional Official Statistics*, *2022*(1), 405–414. https://doi.org/10.34123/semnasoffstat.v2022i1.1480

Mauludin, M. R., Nurdiawan, O., & Basysyar, F. M. (2025). Penerapan Algoritma K-Means Clustering Untuk Analisis Kinerja Pengiriman Paket Shopee Express Di Hub Transit Kedawung. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(1). https://doi.org/10.23960/jitet.v13i1.5870

Putri, D. A., & Gusmanely, Z. (2022). Comparison of K - Means and K - Medoids Algorithms in Clustering Islamic Commercial Banks in Indonesia Using Davies - Bouldin Index , Calinski - Harabasz Index , and Silhouette Coefficient sektor perbankan syariah . Bank (ROA), yang biasanya disebut Ind. *Jurnal Kelitbangan*, *13*(1), 1–12.

Rachman, J. A., & Huda, S. (2025). Faktor-faktor yang Mempengaruhi Penyerapan Tenaga Kerja Sektor Pariwisata di Kabupaten Banyuwangi. *Jurnal Riset Ilmu Ekonomi*, *5*(1), 63–77. https://doi.org/10.23969/jrie.v5i1.249

Rahmawati, T., Wilandari, Y., & Kartikasari, P. (2024). Analisis Perbandingan Silhouette Coefficient Dan Metode Elbow Pada Pengelompokkan Provinsi Di Indonesia Berdasarkan Indikator Ipm Dengan K-Medoids. *Jurnal Gaussian*, *13*(1), 13–24. https://doi.org/10.14710/j.gauss.13.1.13-24

Ramadhan, A., Achmad, F., Zulkarnain, I., & Aritsugi, M. (2025). Evaluation of K-Means, DBSCAN, and Hierarchical Clustering for Strategic Segmentation of Tourism SMEs in Rembang, Indonesia. *Jurnal Teknik Informatika (Jutif)*, *6*(3), 1605–1630. https://doi.org/10.52436/1.jutif.2025.6.3.4602

Wang, D., Lu, X., & Rinaldo, A. (2019). DBSCAN: Optimal rates for density-based cluster estimation. *Journal of Machine Learning Research*, *20*, 1–50.

Yolandari, N. A., & Lastri Elisabet Butarbutar. (2025). ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3). https://doi.org/10.23960/jitet.v13i3.6884