JITET (Jurnal Informatika dan Teknik Elektro Terapan)
Vol. 13 No. 3, pISSN: 2303-0577 eISSN: 2830-7062 http://dx.doi.org/10.23960/jitet.v13i3.6884

ANALISIS PERBANDINGAN K-MEANS DAN DBSCAN DALAM PENGELOMPOKAN DATA TRAVEL REVIEW RATINGS MENGGUNAKAN EVALUASI SILHOUETTE INDEX DAN DAVIES-BOULDIN INDEX

Nezza Anggraini Yolandari[1*], Lastri Elisabet Butarbutar[2], Gloria Citra Hasiana Rajagukguk[3], M. Fikri Zulfi[4], Arnita[5], Fanny Ramadhani[6]

[1,2,3,4,5,6]Universitas Negeri Medan; Jl. Willem Iskandar, Medan, Sumatra Utara

Keywords:
Clustering;
K-Means;
DBSCAN;
Silhouette Index;
Davies-Bouldin Index;

Corespondent Email:
nezzaanggraini0@gmail.com

Abstrak. Dalam era digital, data ulasan wisatawan menjadi sumber informasi penting untuk analisis preferensi dan pengambilan keputusan di sektor pariwisata. Teknik Clustering menjadi salah satu pendekatan yang efektif untuk mengidentifikasi pola tersembunyi dalam data ulasan tersebut. Penelitian ini bertujuan untuk membandingkan performa dua algoritma clustering, yaitu K-Means dan DBSCAN, dalam mengelompokkan data Travel Review Ratings. K-Means menggunakan Elbow Method untuk menentukan jumlah klaster optimal, sedangkan DBSCAN mengandalkan kepadatan data dengan parameter epsilon dan minPts. Evaluasi hasil klasterisasi dilakukan menggunakan metrik Silhouette Index (SI) dan Davies-Bouldin Index (DBI). Hasil penelitian menunjukkan bahwa DBSCAN menghasilkan performa yang lebih baik dibandingkan K-Means dengan nilai SI sebesar 0,27204 dan DBI sebesar 0,83869. DBSCAN dinilai lebih efektif dalam mengidentifikasi struktur klaster yang tidak beraturan serta menangani outlier, sehingga lebih cocok digunakan untuk dataset ulasan wisata yang kompleks.

Abstract. *In the digital era, traveler review data is an important source of information for preference analysis and decision-making in the tourism sector. Clustering technique is one of the effective approaches to identify hidden patterns in the review data. This study aims to compare the performance of two clustering algorithms, namely K-Means and DBSCAN, in clustering Travel Review Ratings data. K-Means uses Elbow Method to determine the optimal number of clusters, while DBSCAN relies on data density with epsilon and minPts parameters. Evaluation of clustering results is done using Silhouette Index (SI) and Davies-Bouldin Index (DBI) metrics. The results show that DBSCAN produces better performance than K-Means with an SI value of 0.27204 and DBI of 0.83869. DBSCAN is more effective in identifying irregular cluster structures and handling outliers, making it more suitable for complex travel review datasets.*

1. PENDAHULUAN

Di era digital saat ini, industri pariwisata mengalami transformasi melalui pemanfaatan teknologi dan data. Salah satu sumber data yang berharga adalah ulasan (review) dari para wisatawan yang memberikan penilaian terhadap destinasi, layanan, maupun pengalaman mereka selama perjalanan. Data ulasan ini, yang sering kali disajikan dalam bentuk rating, mengandung informasi penting yang dapat diolah lebih lanjut untuk mendukung pengambilan keputusan, seperti rekomendasi destinasi, pengembangan layanan, dan segmentasi pasar [1].

Saat ini, jumlah data yang tersedia dari berbagai bidang terus meningkat dengan pesat. Peningkatan ini tidak hanya mempengaruhi cara kita mengumpulkan informasi tetapi juga menuntut metode analisis yang lebih unggul untuk mengelola dan menginterpretasikan data tersebut [2]. Salah satu teknik analisis data yang semakin mendapatkan perhatian adalah Clustering. Clustering merupakan metode dalam unsupervised learning yang bertujuan untuk mengelompokkan data berdasarkan kesamaan karakteristik tertentu [3]. Dalam konteks analisis data *travel review ratings*, teknik clustering dapat digunakan untuk mengelompokkan wisatawan berdasarkan pola penilaian mereka.

Dua algoritma yang populer digunakan dalam proses clustering adalah K-Means dan DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Algoritma K-Means bekerja dengan membagi data ke dalam sejumlah klaster berdasarkan kedekatan terhadap titik pusat (centroid), sementara DBSCAN menggunakan pendekatan berbasis kepadatan untuk menemukan klaster yang memiliki kepadatan tinggi dan mengidentifikasi data pencilan [3].

Namun, efektivitas dari algoritma clustering perlu dievaluasi untuk menentukan hasil pengelompokkan yang optimal. Oleh karena itu, dalam penelitian ini digunakan dua metrik evaluasi yaitu Silhouette Index dan Davies-Bouldin Index. Metode Silhouette Coefficient merupakan metode evaluasi cluster untuk melihat kualitas seberapa baik suatu objek ditempatkan dalam suatu cluster dengan menggabungkan metode cohessian dan separation. Davies-Bouldin Index (DBI) pertama diperkenalkan oleh David L. Davies dan Donald W. Bouldin pada tahun 1979 dengan tujuan untuk mengevaluasi hasil dan untuk menentukan jumlah cluster yang paling optimal dalam proses clustering [4]. Silhouette Index mengukur seberapa mirip suatu objek dengan cluster-nya sendiri dibandingkan dengan cluster lain, sementara Davies-Bouldin Index menilai kualitas cluster berdasarkan rata-rata jarak antar cluster dan dispersi pada cluster.

Melalui penelitian ini, akan dilakukan analisis perbandingan antara algoritma K-Means dan DBSCAN dalam mengelompokkkan data travel review ratings, serta mengevaluasi kualitas hasil pengelompokkannya menggunakan kedua metrik tersebut. Hasil dari penelitian ini diharapkan dapat memberikan wawasan mengenai algoritma clustering yang lebih efektif dalam konteks analisis data pariwisata berbasis rating.

2. TINJAUAN PUSTAKA

2.1 Data Mining

Data mining adalah proses yang memanfaatkan teknik statistik, kecerdasan buatan, dan machine learning untuk mengekstraksi pengetahuan atau informasi tersembunyi dari kumpulan data dalam jumlah besar. Data mining merupakan proses iteratif yang secara otomatis dapat menemukan pola dan hubungan tersembunyi dalam data dengan tujuan memberikan indikasi atau prediksi yang berguna untuk pengambilan keputusan. Data mining tidak hanya memproses data yang sangat besar (big data) namun juga harus mampu memberikan hasil berupa pengetahuan yang mudah dipahami oleh pengguna [5].

2.2 Clustering

Proses menempatkan data atau objek ke dalam kelas atau cluster berdasarkan seberapa mirip atributnya dikenal sebagai clustering. Salah satu metode untuk data mining adalah clustering. Pengelompokan yang baik menghasilkan objek-objek dengan kesamaan rendah dengan yang ada di cluster lain tetapi kesamaan tinggi dengan yang ada di grup atau cluster yang sama [6]. Clustering adalah proses pengelompokan suatu pola yang belum memiliki label dan dilakukan tanpa supervisi menjadi sebuah kelompok yang memiliki karakteristik tertentu. Clustering sangat penting pada beberapa permasalahan antar lain analisa pola, pembuatan keputusan, machine learning, data mining dan sebagainya. [7].

2.3 K-Means

K-Means adalah salah satu algoritma clustering berbasis partisi yang pertama kali diperkenalkan oleh MacQueen (1967). Algoritma ini bekerja dengan mengelompokkan data ke dalam k kelompok berdasarkan jarak Euclidean terhadap centroid yang dihasilkan. Kelebihan utama dari K-Means adalah efisiensinya dalam menangani dataset besar. Namun, kelemahannya adalah ketergantungan pada pemilihan nilai k yang optimal dan

sensitivitas terhadap titik awal centroid. Untuk mengatasi permasalahan ini, metode seperti Elbow Method dan Gap Statistic sering digunakan dalam menentukan jumlah cluster yang optimal [3].

Algoritma K-Means merupakan salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster atau kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai karakteristik berbeda dikelompokkan ke dalam kelompok yang lainnya [8].

2.4 *DBSCAN (Density-Based Spatial* Clustering Algorithm with Noise)

DBSCAN merupakan algoritma pengelompokan yang baik dalam mengatasi outlier atau noise. DBSCAN menghasilkan klaster yang lebih akurat dan baik pada jumlah data besar dan tidak perlu ditentukan jumlah klaster awalnya. Kepadatan data pada DBSCAN dapat identik dengan inti (core), batas (border), serta gangguan (noise) [9].

DBSCAN merupakan algoritma klasterisasi yang tidak memerlukan penentuan jumlah klaster di awal, tetapi bergantung pada dua parameter utama, yaitu Epsilon (ε) dan MinPts. Parameter ε merupakan jarak maksimum antara dua titik agar dapat dianggap berada dalam satu klaster, sedangkan MinPts adalah jumlah minimum titik yang harus berada dalam radius ε untuk membentuk suatu klaster. Dalam proses pengelompokan data, DBSCAN bekerja berdasarkan kepadatan data tanpa menggunakan titik pusat (centroid). Langkah pertama adalah mengidentifikasi titik inti (core point), yakni titik yang memiliki setidaknya MinPts tetangga dalam radius ε. Selanjutnya, titik border (border point) adalah titik yang tidak memiliki jumlah tetangga cukup untuk menjadi titik inti, tetapi masih berada dalam jangkauan ε dari titik inti. Sedangkan titik yang tidak memenuhi kedua kriteria tersebut sebagai outlier (noise point), dan tidak dimasukkan ke dalam klaster manapun [10].

2.5 Silhouette Index (Silhouette Coefficient)

Silhouette Coefficient adalah penggabungan dari metode cohessian dan separation yang merupakan metode evaluasi untuk Cluster. Jarak antara data dapat dihitung menggunakan rumus Euclidean Distance dan Manhattan Distance. Untuk mendapatkan informasi tentang kualitas hasil Clustering pada proses Clustering dapat dihitung menggunakan silhouetter dari masing-masing Cluster bahkan Cluster dari hasil kerja algoritma Clustering [11]. Pada metode ini terdapat beberapa tahapan yaitu sebagai berikut:

Menghitung rata-rata jarak dari suatu dokumen misalkan I dengan semua dokumen lain yang berada dalam satu Cluster

$$a(i) = \frac{1}{|A|-1} \sum j \in A, j \neq i \; d(i,j) \quad (1)$$

Dengan j adalah dokumen lain dalam satu Cluster A dan d (i, j) adalah jarak antara dokumen i dan j. Hitung rata-rata jarak dari dokumen I tersebut dengan semua dokumen di Cluster lain dan diambil nilai terkecilnya.

$$d(i,C) = \frac{1}{|A|} \sum j \in C \; d(i,j) \quad (2)$$

Dengan d (i, C) adalah jarak rata-rata dokumen i dengan semua objek pada Cluster lain C dimana A ≠ C.

$$b(i) = min \; C \neq A \; d(i,C) \quad (3)$$

Nilai Silhouette Coefficientnya adalah:

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

2.6 Davies-Bouldin Index

Davies-Bouldin Index adalah metrik evaluasi yang digunakan untuk mengukur kualitas cluster dalam algoritma K-Means. Nilai DBI dihitung berdasarkan rasio antara jarak antar cluster dan kepadatan intra-cluster. Semakin kecil nilai DBI, semakin baik kualitas cluster [12].

Davies-Bouldin Index (DBI) adalah metrik evaluasi internal yang digunakan untuk menilai kulitas hasil clustering. Metrik ini mengukur seberapa baik setiap klaster terpisah satu sama lain dengan mengevaluasi parameter-parameter seperti kepadataan dan jarak antar klaster. Semakin rendah nilai DBI, semakin baik kualitas clustering, karena menunjukkan bahwa klaster-klaster tersebut lebih terpisah dengan jelas dan tidak saling tumpang tindih. Sebaliknya, nilai DBI yang tinggi mengindikasikan adanya klaster yang kurang terpisah dengan baik, menunjukkan clustering yang kurang efektif [12].

Menurut [13] Langkah pertama ialah mencari rasio atau perbandingan antara rata-rata jarak dalam cluster (Sum of Square Within)

dengan rata-rata jarak antar-cluster (Sum of Square Between) yang dihitung dengan menggunakan Euclidean Distance. Rasio ini didapatkan melalui rumus berikut:

$$R_{ij} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (5)$$

Selanjutnya nilai rasio tersebut digunakan untuk dalam persamaan berikut untuk mendapatkan nilai DBI:

$$DBI = \frac{1}{k} \Sigma_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad (6)$$

2.7 PCA (Principal Componen Analisys)

Principal Componen Analisys (PCA) adalah suatu teknik statistik multivariant yang secara linear mengubah bentuk sekumpulan variabel asli menjadi kumpulan variabel yang lebih kecil yang tidak berkorelasi yang dapat mewakili informasi dari sekumpulan variabel asli. Tujuan utamanya ialah menjelaskan sebanyak mungkin jumlah varian data asli dengan sedikit mungkin komponen utama yang disebut faktor [10].

2.8 Robust Scaler

Robust Scaler dapat mengurangi dampak outlier dan menyelaraskan rentang data tanpa mengasumsikan distribusi normal. Teknik ini sangat efektif untuk data yang sering mengandung nilai ekstrem, seperti yang sering ditemukan dalam dataset ekologis. Robust Scaler bekerja dengan menyesuaikan skala data berdasarkan nilai interquartile range (IQR), sehingga data dengan distribusi yang lebih bervariasi tetap dapat distandardisasi tanpa terpengaruh oleh outlier [14].

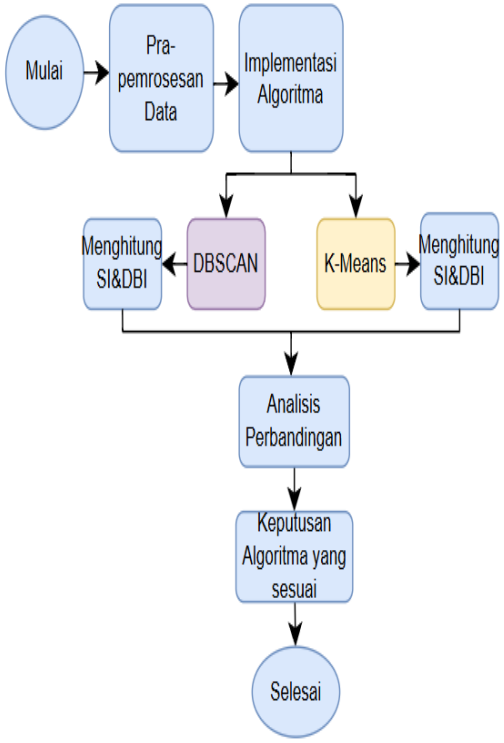
2.9 Interquartile Range (IQR)

Pemeriksaan outlier dilakukan untuk mendeteksi nilai-nilai ekstrem yang berada di luar pola umum data. Outlier dapat memengaruhi distribusi dan akurasi model, terutama pada model-model yang sensitif terhadap skala data. Deteksi outlier dilakukan menggunakan metode *Interquartile Range* (IQR), dengan cara mengidentifikasi data yang berada di bawah Q1 - 1.5×IQR atau di atas Q3 + 1.5×IQR. Nilai- nilai di luar rentang batas bawah dan batas atas dianggap sebagai outlier. *Interquartile Range* (IQR) adalah metode statistika deskriptif untuk mengukur sebaran data di sekitar median. IQR membantu mengidentifikasi keragaman data tanpa terpengaruh oleh nilai ekstrem (outlier). Semakin besar nilai IQR, semakin tinggi variasi data [15].

3. METODE PENELITIAN

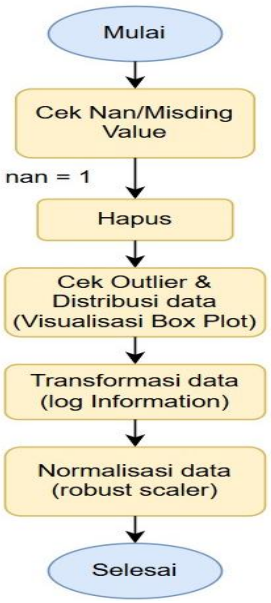
3.1 Pengumpulan Data

Penelitian menggunakan pendekatan kuantitatif dengan metode eksperimen komputasi untuk menganalisis dan membandingkan performa algoritma *clustering* K-Means dan DBSCAN dalam pengelompokan data ulasan (review) perjalanan berdasarkan rating. Data yang digunakan berupa data sekunder yaitu dataset *Travel Review Ratings* yang diperoleh melalui situs https://archive.ics.uci.edu/dataset/485/tarvel+review+ratings. Kumpulan data ini diisi dengan mengambil peringkat pengguna dari ulasan Google. Peringkat pengguna Google berkisar dari 1 hingga 5 dan peringkat pengguna rata-rata per kategori dihitung.



Gambar 1. Alur Penelitian

3.2 Pra-Pemrosesan Data



Gambar 2. Pra-pemrosesan data

1. Tahap pra-pemrosesan dimulai dengan pengecekan data kosong (NaN / missing value).
2. Jika ditemukan, dan jumlahnya sangat kecil (misalnya hanya 1 data), maka data tersebut langsung dihapus.
3. Pengecekan outlier dan distribusi data menggunakan box plot, yaitu visualisasi dari distribusi data numerik yang merangkum lima ukuran statistik utama yaitu, minimum, kuartil pertama (Q1), median (Q2), kuartil ketiga (Q3), dan maksimum. Boxplot digunakan untuk menunjukkan persebaran, kecendrungan sentral, dan potensi pencilan (outliers) dalam suatu dataset [16].
4. Transformasi data menggunakan log information untuk mengatasi distribusi yang skewed. Log Information adalah teknik mengubah data dengan distribusi yang sangat miring (skewed) menjadi lebih simetris. Teknik ini sering diterapkan pada data yang memiliki rentang nilai sangat besar atau data dengan outlier ekstrem.
5. Kemudian, data dinormaliasi menggunakan robust scaler, yaitu metode normalisasi yang tahan terhadap outlier karena menggunakan median dan interkuartil range (IQR). Tahapan ini bertujuan untuk memastikan kualitas data yang optimal sebelum proses klasterisasi dilakukan.

3.3 Implementasi Algoritma

3.3.1 K-MEANS

Berikut tahapan-tahapan dari algoritma K-Means sebagai berikut:

1. Tentukan jumlah klaster (K), menggunakan Elbow Method, yaitu metode menetapkan jumlah (k) ataupun cluster yang akurat pada dataset penelitian. Proses metode elbow mengidentifikasi persentase hasil perbandingan dari jumlah total (k) serta turut menyajikan sebuah lekukan pada grafik akan dinamai siku pada titik [17].
2. Tentukan titik pusat (centroid) dari masing-masing klaster secara acak.
3. Pada setiap titik dihitung jarak terdekat terhadap centroid menggunakan rumus Euclidean Distance.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (7)$$

i = Banyak data
(x,y) = Titik data
(s,t) = Titik pusat

4. Terbentuk klaster baru dengan mengelompokkan setiap titik data ke klaster terdekat berdasarkan jarak Euclidean.
5. Tentukan titik pusat (centroid) baru dengan menghitung rata-rata posisi titik data dalam klister

$$\bar{v} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \quad (8)$$

$\bar{v}_{ij}$ =Rata-rata titik pusat pada cluster ke-i untuk variabel ke-j
N_i =Banyaknya suatu anggota cluster ke-i
i,k = indikator dari cluster
j = indikator dari variabel
x_{kj} =Nilai data ke-k pada cluster tersebut untuk variabel ke-j 6

6. Ulangi tahap 3, 4 dan 5 sampai anggota klister tidak beralih ke klister lainnya maka iterasi berhenti.

3.3.2 DB-SCAN

Berikut tahapan-tahapan dari algoritma DBSCAN sebagai berikut:

1. Tentukan nilai epsilon (ε) dan MinPts. Untuk menentukan nilai ε digunakan pendekatan grafik k-distance, yaitu dengan menghitung jarak ke-k terdekat dari setiap titik data, mengurutkan nilai-nilai jarak tersebut secara menurun, lalu memvisualisasikannya dalam bentuk grafik. Titik di mana kurva mulai menurun (elbow) menjadi acuan untuk memilih nilai ε yang optimal (Maulidhia et al, 2025).
2. Tentukan nilai p atau titik awal secara random atau acak.
3. Menghitung nilai Eps atau hitung jarak masing-masing titik yang memiliki kepadatan terhadap titik p dengan rumus Euclidean Distance berikut:
4. Sebuah cluster terbentuk jika titik sudah mencukupi epsilon lebih dari minimum poin, maka titik tersebut sebagai titik pusat.
5. Ulangi tahap 3 dan 4 sampai semua titik dilakukan perhitungan. Lanjutkan ke titik lainnya ketika tidak terdapat titik yang memiliki kepadatan terhadap p atau titik awal.

3.4 Evaluasi Hasil Klastering

Pada tahap evaluasi hasil klastering, dilakukan pengukuran terhadap kualitas dan performa pengelompokan data menggunakan dua metrik evaluasi, yaitu Silhouette Index (SI) dan Davies-Bouldin Index (DBI) untuk menilai seberapa baik data dikelompokkan oleh algoritma K-Means dan DBSCAN.

1. Hitung dan analisis nilai SI dan DBI kedua algoritma. Nilai SI berada pada rentang -1 hingga 1, di mana semakin tinggi nilainya menunjukkan bahwa data semakin tepat berada dalam klasternya. Nilai DBI semakin rendah maka semakin baik hasil pengelompokan karena menunjukkan pemisahan yang jelas antar klaster.
2. Membandingkan performa algoritma K-Means dan DBSCAN berdasarkan hasil SI dan DBI untuk mengetahui algoritma mana yang lebih optimal dalam mengelompokkan data *travel review ratings*. Penilaian dilakukan berdasarkan nilai SI yang lebih tinggi dan DBI yang lebih rendah sebagai indikator performa klasterisasi yang baik.

4. HASIL DAN PEMBAHASAN

4.1 Persiapan Data

Data yang digunakan pada penelitian merupakan Dataset Travel Review Ratings. Dataset klasifikasi dengan tipe numerik terdiri dari 16 fitur berupa tempat wisata (User ID, Churches, Resorts, Beaches, Parks, Threatres, Museums, Malls, Zoo, Restaurants, Pubs/bars, Local Services, Burger/pizza shops, Hotels, Juice Bars, Art Galleries, Dance Clubs, Swimming Pools, Gyms, Bakeries, Beauty Spas, Cafes, View Points, Monuments, Gardens) dan 5456 data berupa pengguna.

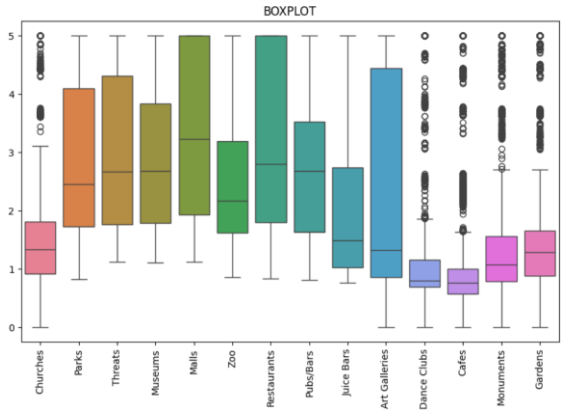
Tabel 1. Dataset

User ID	Churches	Resorts	…	Monuments	Gardens
User 1	0	0	…	0,5	0,53
User 2	0	0,5	…	0,51	0,54
User 3	0	0	…	0,51	0,55
…	…	…	…	…	…
User 5456	0,95	4,07	…	5	1,17

4.2 Pra-Pemrosesan Data

Pada tahap awal pengolahan data, dilakukan pemeriksaan terhadap keberadaan nilai kosong atau NaN (Not a Number) dalam dataset travel review ratings. Hal ini penting agar analisis klaster yang dilakukan menggunakan metode K-Means dan DBSCAN berjalan dengan data yang bersih dan konsisten. Nan ditemukan pada fitur *Gardens* pada user *1347* sebanyak 1. Karna NaN yang ditemukan relatif kecil dan tidak memiliki dampak yang signifikan maka langkah yang dipilih dalam mengatasi NaN adalah penghapusan. Setelah penghapusan Nan terdapat 5455 data.

Untuk memahami karakteristik distribusi data dan mendeteksi keberadaan outlier pada fitur-fitur tempat wisata, dilakukan visualisasi menggunakan boxplot.



Gambar 3. Hasil Boxplot

Berdasarkan Boxplot, hasil distribusi keseluruhan data yaitu Skewness 0,77 menunjukkan bahwa distribusi condong ke kanan (Right Skewed) berarti terdapat lebih banyak nilai-nilai kecil dan sedikit nilai yang sangat tinggi dan Kurtosis -0,684 menunjukkan bahwa distribusi data memiliki puncak lebih datar dibandingkan distribusi normal dan cenderung memiliki penyebaran data yang lebih merata.

Melalui visualisasi box plot, outlier diidentifikasi sebagai nilai yang berada di luar rentang interkuartil (IQR). Dari 16 fitur di dalam dataset terdapat 5 fitur yang dideteksi memiliki jumlah oulier yang cukup tinggi. Total outlier di keseluruhan dataset 2450. Dan 9 fitur lainnya menunjukkan tidak ada outlier.

Tabel 2. Total Outlier

Fitur	Jumlah Outlier
Churches	197
Dance Clubs	501
Cafes	486
Monuments	718
Gardens	548

Hasil ini menunjukkan bahwa beberapa fitur, terutama Monuments, memiliki penyebaran data yang sangat lebar dengan banyak nilai yang berada jauh di luar rentang IQR (interquartile range). Outlier dapat muncul karena adanya perbedaan signifikan dalam preferensi atau popularitas tempat wisata yang jauh lebih tinggi atau rendah dibandingkan tempat lainnya.

Outlier dan distribusi data yang skewed dapat memengaruhi hasil klustering sehingga normalisasi atau transformasi data diperlukan untuk memastikan klusterisasi lebih optimal. Penanganan yang dilakukan dengan transformasi data yaitu log information dan normalisasi data yaitu robust scaler. Beberapa contoh data setelah transformasi dan normalisasi:

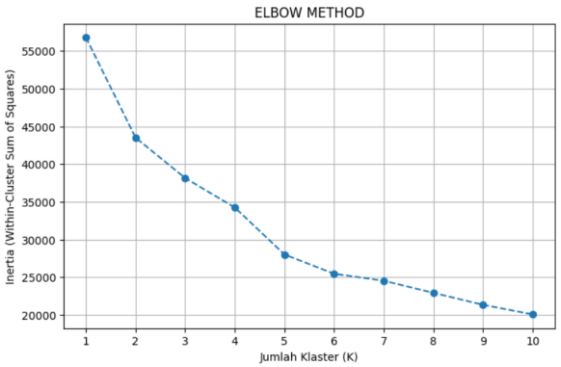
	Churches	Parks	Threats	Museums	Malls	Zoo	Restaurants	Pubs/Bars	Juice Bars
0	-2.232192	0.473748	0.755393	0.114904	0.487693	0.117627	-0.173234	-0.020283	0.144586
1	-2.232192	0.473748	0.755393	0.114904	0.487693	0.294450	-0.173234	-0.015191	0.144586
2	-2.232192	0.466840	0.755393	0.114904	0.487693	0.294450	-0.173234	-0.020283	0.144586
3	-2.232192	0.466840	0.755393	0.114904	0.487693	0.117627	-0.173234	-0.020283	0.144586
4	-2.232192	0.466840	0.755393	0.114904	0.487693	0.294450	-0.173234	-0.020283	0.144586

(a)

Gambar 4. Contoh data setelah transformasi

4.3 Algoritma K-Means

Langkah awal dalam implementasi algoritma K-Mean adalah menentukan jumlah klaster optimal yang akan digunakan. Menggunakan Elbow Method, yang merupakan metode visual berbasis metrik inertia atau sum of squared errors (SSE) terhadap jumlah klaster yang berbeda-beda. Berikut adalah hasil Elbow Method:



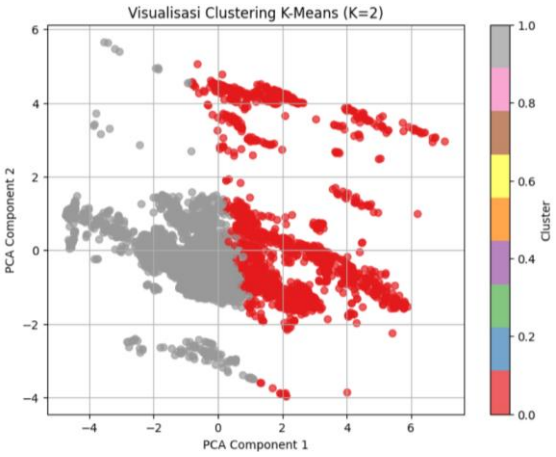
Gambar 5. Hasil Elbow Method

Nilai inertia menunjukkan seberapa baik data dikelompokkan maka semakin kecil nilainya, berarti jarak antar titik dalam klaster semakin kecil dan pengelompokan semakin baik. Saat nilai K meningkat dari 1 ke 2, terjadi penurunan inertia yang cukup tajam. Namun, setelah K = 2, penurunan nilai inertia mulai melambat. Pola ini membentuk suatu titik siku (elbow) pada grafik. Titik inilah yang menunjukkan jumlah klaster optimal, karena setelah titik tersebut, penambahan klaster tidak lagi memberikan peningkatan yang signifikan dalam kualitas pengelompokan. Titik siku pada

grafik terjadi di K=2, sehingga dipilihlah jumlah klaster optimal sebanyak 2 klaster. Dengan terbentuknya dua kluster, dapat diasumsikan bahwa data review rating tempat wisata cenderung terbagi menjadi dua kelompok utama. Klaster pertama kemungkinan besar berisi tempat-tempat wisata yang mendapatkan rating tinggi dari para pengguna. Dan klaster kedua mungkin terdiri dari tempat-tempat wisata yang kurang disukai atau mendapat rating rendah.

Jumlah kluster optimal telah ditentukan sebesar K=2, maka algoritma memilih dua titik awal sebagai centroid awal kluster 0 dan kluster 1. Setiap data dihitung jaraknya terhadap masing masing centroid menggunakan Euclidean Distance. Data akan dikelompokkan ke centroid yang jaraknya paling dekat. Setelah semua data dikelompokkan, centroid dari masing-masing kluster dihitung ulang sebagai rata-rata dari semua data dalam kluster tersebut. Proses diulangi sampai proses centroid tidak berubah (konvergen).

Jumlah data pada kluster 0 = 1809 data dan kluster 1 = 3646 data. Berdasarkan hasil analisis posisi centroid, terdapat perbedaan karakteristik antara kluster 0 dan 1 dalam hal preferensi jenis tempat wisata.



Gambar 6. Grafik hasil klusterisasi K-Means

Grafik diatas adalah visualisasi hasil klusterisasi yang menampilkan hasil reduksi menggunakan PCA (Principal Components Analysis) untuk memproyeksikan data berdimensi tinggi ke dalam 2 dimensi utama (PCA component 1 dan 2) sehingga distribusi kluster bisa divisualisasikan dengan jelas.

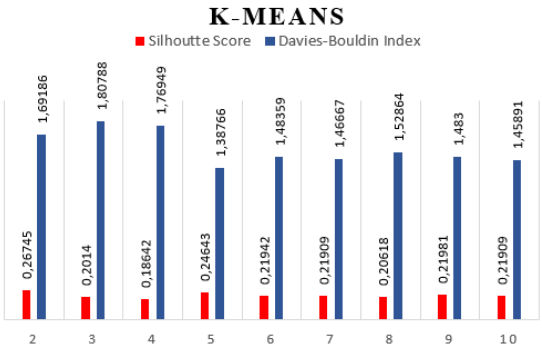
Evaluasi kualitas kluster menggunakan Silhouette Index dan Davies-Bouldin Index.

Tabel 3. Evaluasi Kualitas Kluster K-Means

K-Means		
Kluster	Silhoutte Score	Davies-Bouldin Index
2	0,26745	1,69186
3	0,2014	1,80788
4	0,18642	1,76949
5	0,24643	1,38766
6	0,21942	1,48359
7	0,21909	1,46667
8	0,20618	1,52864
9	0,21981	1,483
10	0,21909	1,45891

Pada K=2 merupakan hasil dari optimal k elbow method, memiliki nilai SI (0,26745) tertinggi yang menunjukkan pemisahan antar kluster paling baik dibanding lainnya. Sedangkan nilai DBI (1,69186) cukup tinggi yang menunjukkan bahwa ada kemiripan antar kluster atau belum ideal. Pada K=5, nilai SI (0,24643) lebih rendah dibandingkan K=2, namun memiliki nilai DBI (1,38766) terendah yang menunjukkan kluster paling terpisah satu sama lain. Pada nilai K lainnya, nilai SI konsisten di rata-rata 0,2 dan DBI lebih tinggi dibanding K=5.

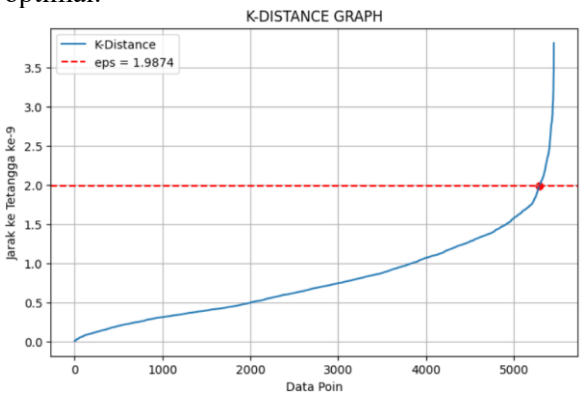
Berdasarkan hasil tersebut, K=2 adalah hasil dari elbow method yang mempertimbangkan within-kluster inertia bukan kualitas pengelompokkan antar kluster. Dari hasil evaluasi SI dan DBI, K=5 memberikan hasil terbaik diantara K lainnya yaitu nilai SI cukup tinggi dan DBI terendah. Sehingga K=5 lebih optimal dibandingkan K=2 dalam hal kualitas klustering meskipun bukan hasil optimal dari elbow method.



Gambar 7. Hasil K-Means

4.4 Algoritma DBSCAN

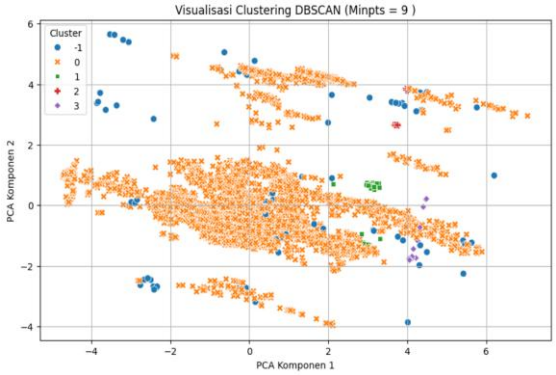
Sebelum menerapkan algoritma DBSCAN, penting untuk menentukan dua parameter utama yang sangat mempengaruhi hasil klustering, yaitu epsilon (eps) dan minimun points (minpts). Minpts umumnya berdasarkan heuristik, lebih besar atau sama dengan dimensi data. Untuk dataset travel review ratings, nilai minpts = 9 cukup optimal. Setelah minpts ditentukan, dapat menentukan nilai eps menggunakan metode K-distance graph. Menghitung jarak tetangga ke-minPts terdekat untuk setiap titik. Mencari titik tekuk dari grafik yang menunjukkan perubahan tajam dalam kemiringan dianggap sebagai nilai eps yang optimal.



Gambar 8. Grafik K-distance Graph

Berdasarkan grafik, dengan nilai minpts = 9 terlihat bahwa titik tekuk terjadi di sekitar jarak 1,9874 yang ditetapkan sebagai nilai eps. Grafik merah horizontal menunjukkan nilai eps = 1.9874, yaitu batas jarak maksimum antara titik yang dianggap berada di satu kluster.

Algoritma DBSCAN akan memulai dengan memilih sembarang titik data yang belum dikunjungi. Titik akan diuji dengan memiliki minimal 9 tetangga dalam radius eps = 1.9874. Jika titik tersebut memenuhi syarat sebagai core point, maka titik-titik tetangganya akan dimasukkan ke dalam satu kluster yang sama, dan pencarian akan berlanjut ke tetangga-tetangganya. Proses ini dilakukan secara iteratif dan rekursif sehingga terbentuk wilayah yang padat (dense). Titik-titik yang tidak memiliki cukup tetangga akan diperiksa ulang. Jika berada di sekitar core point, akan ditandai sebagai border point dan dimasukkan ke kluster terdekat. Jika tidak termasuk kluster manapun dianggap sebagai noise. Diperoleh 4 kluster (tidak termasuk noise) dengan jumlah data pada kluster 0 = 5300 data, kluster 1 = 45 data, kluster 2 = 10 data, dan kluster 3 = 16 data.



Gambar 9. Grafik hasil klustering DBSCAN

Grafik menunjukkan hasil penyebaran hasil klustering dalam dua komponen utama PCA. Kluster 0 sangat dominan dan tersebar padat ditengah, sebagian besar data memiliki kepadatan yang cukup tinggi dan saling berdekatan. Kluster -1 atau noise tersebar di pinggir dan terisolasi menunjukkan bahwa ada data yang berada di luar radius eps dari tetangga terdekatnya. Sedangkan kluster kecil lainnya berukuran jauh lebih kecil dan terbentuk dari kelompok-kelompok data yang padat namun terpisah dari kluster utama.

Evaluasi kualitas kluster menggunakan Silhouette Index dan Davies-Bouldin Index

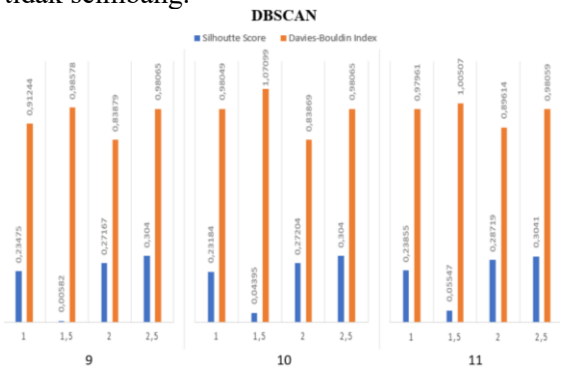
Tabel 4. Evaluasi Kualitas Kluster DBSCAN

DBSCAN				
Min pts	Epsilon	Jumlah Kluster	Silhoutte Score	Davies-Bouldin Index
9	1	73	0,23475	0,91244
	1,5	22	0,00582	0,98578
	2	4	0,27167	0,83879
	2,5	3	0,304	0,98065
	1	66	0,23184	0,98049

10	1,5	20	0,04395	1,07099
	2	4	0,27204	0,83869
	2,5	3	0,304	0,98065
11	1	59	0,23855	0,97961
	1,5	18	0,05547	1,00507
	2	3	0,28719	0,89614
	2,5	3	0,3041	0,98059

Pada minpts=9 dan eps=2 dari hasil K-distance graph, memiliki nilai SI = 0,27167 dan DBI = 0,83879 membentuk 4 kluster. Nilai SI terbaik pada minpts=11 dan eps=2,5 yaitu 0,3041 tetapi nilai DBI (0,98059) relatif tinggi dan membentuk 3 kluster. Nilai DBI terbaik pada minpts=10 dan eps=2 yaitu 0,83869 dengan nilai SI (0,27204) cukup baik membentuk 4 kluster. Nilai eps=2 pada minpts 9 dan 10 memiliki performa yang sama baik untuk kedua metrik evaluasi. Sedangkan nilai eps 1 dan 1,5 pada minpts manapun tidak memberikan pengaruh yang signifikan.

Pengujian eps=2 pada minpts=9 sudah cukup optimal. Namun, pada minpts=10 memberikan nilai DBI sedikit lebih baik tanpa menurunkan nilai SI. Sedangkan pada minpts=11 dan eps=2,5 memberikan nilai SI terbaik tetapi nilai DBI memburuk sehingga tidak seimbang.



Gambar 10. Hasil DBSCAN

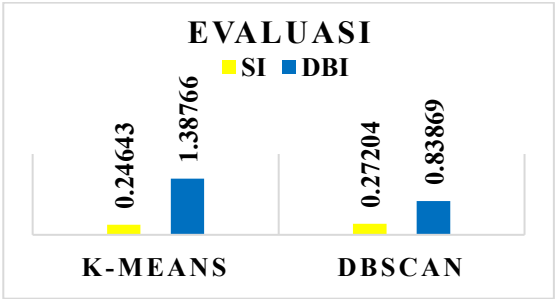
4.5 Evaluasi Klusterisasi

Berikut adalah pembahasan untuk hasil evaluasi klaster dari algoritma K-Means dan DBSCAN berdasarkan metrik Silhouette Score (SI) dan Davies-Bouldin Index (DBI):

Tabel 5. Evaluasi Klaster Pada K-Means & DBSCAN

Algoritma	Kluster	SI	DBI
K-Means	5	0,24643	1,38766
DBSCAN	4	0,27204	0,83869

Algoritma DBSCAN menghasilkan nilai SI lebih tinggi dibandingkan K-Means. Kluster yang dibentuk oleh DBSCAN lebih kompak dan terpisah dengan lebih baik, sehingga kualitas pengelompokkan lebih baik. Nilai DBI pada DBSCAN lebih rendah dari pada K-Means yang berarti lebih terpisah dan kurang tumpang tindih. Hasil DBSCAN dinilai lebih baik dalam memisahkan Kluster dibandingkan K-Means. K-means membentuk 5 kluster sementara DBSCAN menghasilkan 4 kluster.



Gambar 11. Perbandingan SI & DBI Algoritma K-Means

Algoritma DBSCAN memberikan hasil klasterisasi yang lebih optimal dibandingkan K-Means, dengan nilai SI yang lebih tinggi dan DBI yang lebih rendah. Hal ini menunjukkan bahwa DBSCAN lebih cocok digunakan pada dataset travel review ratings. DBSCAN tidak memerlukan input jumlah klaster terlebih dahulu, sehingga mampu menyesuaikan jumlah klaster secara otomatis berdasarkan kepadatan data. Sedangkan K-Means harus menententukan jumlah klasternya di awal, dan kemungkinan tidak optimal jika jumlah klaster yang dipilih tidak sesuai dengan struktur dataset.

5. KESIMPULAN

Berdasarkan hasil pengujian dan evaluasi terhadap algoritma K-Means dan DBSCAN dalam pengelompokkan data *Travel Review Ratings*, dapat disimpulkan bahwa:

a. Algoritma DBSCAN menghasilkan performa klasterisasi yang lebih baik dibandingkan K-Means berdasarkan dua metrik evaluasi yang digunakan, yaitu: Silhouette Index (SI) yang lebih tinggi (0,27204). Davies-Bouldin Index (DBI) yang lebih rendah (0,83869)
b. DBSCAN lebih unggul karena mampu mengelompokkan data secara otomatis

berdasarkan kepadatan, tanpa perlu menentukan jumlah klaster di awal, serta lebih efektif dalam mengidentifikasi outlier dan noise.

c. K-Means cenderung kurang optimal pada dataset yang memiliki outlier, distribusi tidak normal, dan tidak memiliki bentuk klaster yang bulat atau homogen, karena keterbatasannya yang membutuhkan jumlah klaster ditentukan sebelumnya dan kepekaan terhadap penyimpangan data.

d. Pada dataset travel review ratings yang bersifat padat, tidak seimbang, dan mengandung outlier, algoritma DBSCAN lebih cocok digunakan untuk menghasilkan pengelompokkan yang valid dan bermakna.

UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada berbagai pihak yang memberikan kontribusi dan dukungan dalam proses penelitian ini. Secara khusus apresiasi ditujukan kepada Ibu Arnita dan Ibu Fanny Ramadhani yang berperan sebagai pembimbing, serta rekan-rekan yang telah memberikan bantuan selama kegiatan penelitian dan penulisan berlangsung.

DAFTAR PUSTAKA

[1] F. Fahira and C. Prianto, "Prediksi Pola Kedatangan Turis Mancanegara dan Menganalisis Ulasan Tripadvisor dengan LSTM dan LDA," *J. Tekno Insentif*, vol. 17, no. 2, pp. 69–83, 2023.

[2] S. Mutiah, Y. Hasnataeni, A. Fitrianto, and L. M. R. D. Jumansyah, "Perbandingan Metode Klastering K-Means dan DBSCAN dalam Identifikasi Kelompok Rumah Tangga Berdasarkan Fasilitas Sosial Ekonomi di Jawa Barat Dalam era digital saat ini , jumlah data yang tersedia dari berbagai bidang , termasuk sosial dan ekonomi , terus," vol. 09, no. September, pp. 247–260, 2024.

[3] A. Nur, A. Maulidhia, I. Ika, W. Friska, I. Sukarno, and R. Basya, "Implementasi Perbandingan Algoritma k-Means dan DB-Scan Pada Beban Listrik Rumah Tangga," pp. 85–94, 2021.

[4] A. Tania, T. Handhayani, and J. Hendryli, "Perbandingan Antara Algoritma K-Means Dan Algoritma Bisecting K-Means Dalam Menganalisis Gempa Bumi Di Indonesia," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 8, no. 2, pp. 265–270, 2023.

[5] A. A. Baskara, N. M. Piranti, and M. F. Romdendine, "Framework Data Mining : Sebuah Survei," vol. 9, no. 3, pp. 4886–4895, 2025.

[6] I. Pii, N. Suarna, and N. Rahaningsih, "Penerapan Data Mining Pada Penjualan Produk Pakaian Dameyra Fashion Menggunakan Metode K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 423–430, 2023.

[7] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021.

[8] A. A. Zulyani, A. S. Y. Irawan, and A. Jamaludin, "Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Menentukan Tingkat Vaksinasi Pada Kecamatan Tambun Selatan," *J. Soc. Sci. Res.*, vol. 3, no. 3, pp. 7037–7050, 2023.

[9] G. A. Rahman *et al.*, "Dalam Pengelompokan Kabupaten / Kota Di Sulawesi Tenggara Berdasarkan Indikator," vol. 10, no. 1, pp. 184–193, 2025.

[10] M. Wangge, "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA," *J. Cendekia J. Pendidik. Mat.*, vol. 5, no. 2, pp. 974–988, 2021.

[11] W. Wahyu Pribadi, A. Yunus, and A. S. Wiguna, "Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 493–500, 2022.

[12] M. R. Mauludin, O. Nurdiawan, and F. M. Basysyar, "Penerapan Algoritma K-Means Clustering Untuk Analisis Kinerja Pengiriman Paket Shopee Express Di Hub Transit Kedawung," vol. 13, no. 1, pp. 1188–1192, 2025.

[13] L. B. A. 2025 Prasetya, "Computer Based Information System Journal Clustering Dalam Menentukan Tindak Lanjut Hasil Annual Check Mental Health Dengan Algoritma K-Lorensius Bima Ade Prasetya," vol. 01, pp. 55–61, 2025.

[14] R. S. Gumelar, M. Akrom, and G. A. Trisnapradika, "Optimasi model machine learning untuk prediksi inhibitor korosi berbasis augmentasi dataset senyawa n-heterocyclic menggunakan KDE Machine learning model optimization for corrosion inhibitor prediction based on n-heterocyclic compound dataset augmentation using KDE," vol. 10, no. 1, pp. 1–12, 2025.

[15] R. Efendi, A. Junaidi, and A. M. Rizki, "Penentuan Pusat Klaster Secara Otomatis

Pada Algoritma Density Peaks Clustering Berbasis Metode Inter Quartile Range," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, 2024.

[16] A. Deli, P. K. Kondang, W. D. Awil, and A. Ranti, "Analisis Segmentasi Anggaran Pemasaran dan Penjualan Produk di Industri Retail Menggunakan K-Means Clustering Berbasis R Shiny," vol. 4, no. 1, pp. 41–54, 2025.

[17] D. Setiadi *et al.*, "Penerapan Algoritma K-Means Clustering Pada Pembesaran," vol. 7, no. 6, pp. 3320–3327, 2023.