SEMNAS
PUBLIKASI HASIL PENELITIAN DAN PENGABMAS
UNIVERSITAS MUHAMMADIYAH SEMARANG

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Perbandingan Metode K-Means dan K-Medoids dalam Klasterisasi Pencemaran Lingkungan Studi Kasus: Berdasarkan Jumlah Desa Provinsi di Indonesia

Comparison of K-Means and K-Medoids Methods in Environmental Pollution Clustering (Case Study: Based on the Number of Villages per Province in Indonesia)

Nasyiatul Izzah[1], Al Aghni Naufalia[1], Rachmat Kahfiwan Nur[1], Gita Rahmawati[1], Rahayu Setyaningsih[1], Arafico Mirah Delima Dinirvana[1], Fatkhurokhman Fauzi[1]

[1] Universitas Muhammadiyah Semarang, Kota Semarang

Corresponding author : : *nasyiatulizzah04@gmail.com*

Abstrak

Latar belakang: Pencemaran lingkungan, seperti pencemaran air, tanah, dan udara, masih menjadi persoalan utama di banyak wilayah Indonesia. Meskipun data dari Badan Pusat Statistik (BPS) menunjukkan banyak desa dan kelurahan terdampak, belum tersedia segmentasi provinsi yang komprehensif untuk mendukung kebijakan berbasis data. Penelitian ini memfokuskan pada pengelompokan 38 provinsi di Indonesia berdasarkan jenis pencemaran lingkungan menggunakan dua metode klasterisasi yaitu K-Means dan K-Medoids. Tujuannya adalah mengevaluasi metode mana yang lebih optimal dalam membentuk cluster wilayah terdampak. **Metode:** Penilaian dilakukan dengan Davies-Bouldin Index untuk menentukan jumlah cluster. **Hasil Penelitian:** terbaik. Hasil menunjukkan bahwa dua cluster merupakan konfigurasi optimal, dengan K-Means menunjukkan performa lebih baik (DBI = 0.432) dibandingkan K-Medoids (DBI = 0.465). **Kesimpulan:** Oleh karena itu, K-Means direkomendasikan sebagai metode yang lebih efektif dalam pengelompokan wilayah berdasarkan tingkat pencemaran dengan karakteristik cluster menunjukkan cluster 1 mewakili wilayah dengan pencemaran rendah dan cluster 2 mewakili wilayah dengan pencemaran tinggi.

Kata Kunci : pencemaran lingkungan, klasterisasi, provinsi, k-means, k-medoids

Abstract

Background: *Environmental pollution, such as water, soil, and air pollution, remains a major issue in many regions of Indonesia. Although data from Statistics Indonesia (BPS) indicate that many villages and urban neighborhoods are affected, a comprehensive provincial-level segmentation to support data-driven policymaking is still unavailable. This study focuses on clustering 38 provinces in Indonesia based on types of environmental pollution using two clustering methods, namely K-Means and K-Medoids. The objective is to evaluate which method is more optimal in forming clusters of affected regions.* ***Method:*** *The evaluation was carried out using the Davies-Bouldin Index to determine the optimal number of clusters.* ***Result:*** *The results show that two clusters represent the optimal configuration, with K-Means demonstrating better performance (DBI = 0.432) compared to K-Medoids (DBI = 0.465).* ***Conclusion:*** *Therefore, K-Means is recommended as the more effective method for clustering regions based on pollution levels with the cluster characteristics showing that Cluster 1 represents regions with low pollution levels, while Cluster 2 represents regions with high pollution levels.*

Keywords : *environmental pollution, clustering, provinces, k-means, k-medoids*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

PENDAHULUAN

Lingkungan hidup dapat mencangkup segala hal, termasuk makhluk hidup, objek, dan kondisi yang mencangkup manusia dan perilaku yang berdampak pada kelangsungan hidup makhluk hidup lainnya. Seiring dengan kenaikan kualitas hidup, ada pula peningkatan jumlah populasi dan sumber daya alam yang disalahgunakan untuk mencukupi keperluan manusia. Permasalahan ini juga terjadi di beberapa negara, tidak hanya Indonesia. Akibatnya, banyak persoalan lingkungan yang muncul, seperti pencemaran lingkungan yang mencakup pencemaran air, tanah, dan udara [1]. Pencemaran disebabkan oleh aktivitas seperti sisa-sisa rumah tangga, bahan dari industri, bahan kimia yang berbahaya, kebocoran minyak, dan sebagainya [2]. Berdasarkan informasi dari Badan Pusat Statistik (BPS) tahun 2021, terdapat 10.683 desa atau kelurahan yang terpapar pencemaran air. Selain itu, tercatat 1.499 desa atau kelurahan yang mengalami pencemaran tanah, 5.644 desa atau kelurahan yang terpapar pencemaran udara, dan sebesar 69.966 desa atau kelurahan tidak pernah terpapar pencemaran [3]. Melihat banyaknya desa atau kelurahan yang terkena dampak dari berbagai jenis pencemaran lingkungan, dibutuhkan metode analisis yang dapat mengelompokkan desa atau kelurahan berdasarkan kesamaan jenis pencemarannya.

Dalam penelitian ini, digunakan metode *clustering (unsupervised learning)* yaitu *K-Means* dan *K-Medoids*. Algoritma *K-Means* terkenal sebagai teknik penggelompokan yang banyak digunakan, karena kesederhanaan dan kemudahan dalam penerapannya. Selain itu, *Metode K-Means* dan *K-Medoids* digunakan karena data yang dianalisis bersifat numerik, tidak berlabel, dan bertujuan untuk menemukan pola pengelompokan alami. *K-Means* efektif untuk data multivariat dalam jumlah sedang, sedangkan *K-Medoids* digunakan sebagai pembanding karena lebih tangguh terhadap outlier. Dalam data mining, *K-Means* merupakan salah satu teknik pengelompokan data yang dapat mengategorikan data ke dalam satu atau beberapa kelompok [4]. Sedangkan, algoritma *K-Medoids* adalah salah satu metode dalam analisis cluster yang digunakan untuk mengelompokan data berdasarkan kesamaan titik-titik data. *Medoids* sendiri adalah titik data aktual dalam himpunan data yang paling mewakili cluster secara keseluruhan [5].

Penelitian terdahulu menunjukkan bahwa algoritma *clustering* seperti *K-Means dan K-Medoids* telah banyak digunakan untuk mengategorikan data wilayah berdasarkan indikator lingkungan. Dalam studi oleh Sipayung (2020), metode *K-Means* diterapkan untuk mengategorikan desa atau kelurahan di Indonesia sesuai jenis pencemarannya (air, tanah, dan udara). Hasilnya menunjukkan bahwa data dapat dibagi ke dalam dua cluster, yaitu cluster tinggi yang terdiri dari 4 provinsi dan cluster rendah dengan 30 provinsi. *Cluster* tinggi mencerminkan wilayah dengan tingkat pencemaran yang dominan di ketiga jenis tersebut, dan hasil pengelompokan ini juga didukung secara visual melalui *RapidMiner*, yang menunjukkan konsistensi antara hasil manual dan hasil sistem [6]. Sementara itu, penelitian oleh Anjelita (2020) membandingkan algoritma *K-Means* dan *K-Medoids* dalam kasus pencemaran air berdasarkan data dari

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

34 provinsi. Keduanya menghasilkan dua *cluster*, namun distribusi anggota *cluster* berbeda: *K-Means* menghasilkan 4 provinsi dalam cluster pencemaran tinggi dan 30 dalam *cluster* rendah, sedangkan *K-Medoids* menghasilkan kebalikannya. Evaluasi menggunakan *Cluster Distance Performance* menunjukkan bahwa *K-Means* lebih unggul dengan nilai -59084.803 dibandingkan -66471.853 pada *K-Medoids,* yang mengindikasikan bahwa *K-Means* menghasilkan pemisahan *cluster* yang lebih baik berdasarkan kedekatan antar data dalam masing-masing kelompok [1]. Penelitian oleh Maulana (2023) juga menggunakan algoritma *K-Means* untuk mengategorikan kota/kabupaten di Jawa Tengah menurut jumlah desa terdampak pencemaran. Hasilnya membentuk dua *cluster* dan menunjukkan struktur yang baik, dengan nilai *silhouette* 0,514. *Cluster* rendah mencakup 24 wilayah, sementara *cluster* tinggi terdiri dari 14 wilayah, terutama di bagian barat dan tengah Jawa Tengah. Visualisasi menunjukkan pemisahan yang konsisten, menegaskan efektivitas *K-Means* dalam mengidentifikasi daerah prioritas pencemaran [3].

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk mengelompokkan 38 provinsi di Indonesia berdasarkan tingkat pencemaran lingkungan hidup menurut jumlah desa yang terdampak. Dalam penelitian ini digunakan dua metode klasterisasi, yakni *K-Means* dan *K-Medoids,* guna membandingkan performa keduanya dalam membentuk cluster yang representatif. Hasil dari penelitian ini diharapkan dapat memberikan pemahaman yang lebih jelas mengenai sebaran wilayah dengan tingkat pencemaran lingkungan yang tinggi maupun rendah, serta menjadi dasar pertimbangan bagi pemerintah, masyarakat, dan kalangan akademisi dalam merumuskan kebijakan pengendalian pencemaran lingkungan secara lebih efektif dan terarah.

METODE

Metode dan Sumber Data

Penelitian ini memanfaatkan data sekunder yang diambil dari Badan Pusat Statistik (BPS) tahun 2024. Data tersebut mencakup informasi mengenai jumlah desa/kelurahan berdasarkan jenis pencemaran lingkungan hidup yang dikelompokkan menurut provinsi di Indonesia. Jenis pencemaran yang dimaksud antara lain mencakup pencemaran air, udara, dan tanah.

Data ini digunakan untuk menganalisis sebaran dan karakteristik pencemaran lingkungan hidup di berbagai wilayah di Indonesia, dengan variabel-variabel diuraikan dalam Tabel 1 berikut.

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Tabel 1. Variabel Penelitian

Variabel	Keterangan
X1	Jumlah desa/kelurahan yang terdampak pencemaran air tahun 2024
X2	Jumlah desa/kelurahan yang terdampak pencemaran tanah tahun 2024
X3	Jumlah desa/kelurahan yang terdampak pencemaran udara tahun 2024
X4	Jumlah desa/kelurahan yang tidak mengalami pencemaran lingkungan hidup tahun 2024

Uji Multikolinearitas

Multikolinieritas diuji untuk mengidentifikasi adanya hubungan yang tinggi atau sempurna antar variabel independen. Ketika sebagian atau seluruh variabel bebas dalam model regresi menunjukkan nilai korelasi yang sempurna, maka kondisi tersebut menunjukkan adanya multikolinieritas [7]. Suatu model regresi dikatakan baik apabila tidak ditemukan korelasi sempurna antar variabel bebasnya. Keberadaan multikolinieritas dapat mengakibatkan ketidakakuratan dalam penerapan metode regresi, karena estimasi parameter regresinya menjadi tidak stabil dan nilai koefisien cenderung membesar secara tidak wajar [8]. Untuk mendeteksi Multikolinieritas, umumnya digunakan beberapa metode seperti perhitungan *Variance Inflation Factor* (VIF), analisis korelasi Pearson antar variabel independen, serta pengamatan terhadap nilai Condition Index (CI) dan eigenvalues. Suatu model regresi dapat disimpulkan bebas dari masalah multikolinieritas apabila hasil uji menunjukkan nilai Tolerance > 0.01 dan nilai VIF < 10 [9].

Algoritma *K-Means*

Dalam proses data mining, algoritma *K-Means* digunakan sebagai teknik klastering yang mengelompokkan data berdasarkan jarak antar data (Sekar Setyaningtyas et al., 2022). Metode *K-Means* termasuk dalam kategori klastering non-hierarki yang membagi data menjadi dua atau lebih klaster. Pengelompokan ini dilakukan dengan menyusur data ke dalam beberapa kelompok, di mana data yang memiliki karakteristik yang serupa akan ditempatkan dalam satu klaster, sedangkan data dengan karakteristik yang berbeda akan dipisahkan ke dalam klaster lainnya [10]. Tujuan dari proses pengelompokan ini adalah untuk meminimalkan fungsi objektif dengan cara memperkecil variasi dalam satu kelompok dan memperbesar perbedaan antar kelompok [11]. Adapun tahapan-tahapan dalam Algoritma *K-Means* sebagai berikut [12]:

1. Menetapkan jumlah klaster (k) yang digunakan oleh peneliti sebagai awal proses pengelompokan
2. *Centeroid* awal kemudian dipilih secara acak sebagai pusat awal masing-masing klaster
3. Jarak antara setiap titik data dengan seluruh *Centeroid* dihitung menggunakan rumus *Euclidean Distance,* untuk menentukan kedekatan relatif setiap data

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

terhadap pusat klaster. Persamaan *Euclidean Distance* dirumuskan sebagai berikut:

$$D(x,z) = \sqrt{\sum_{i=1}^{n}(x_i - z_i)^2} \quad (1)$$

di mana:

- $D(x,z)$ adalah jarak *Euclidean* antara dua titik (x,z)
- x_i, z_i merupakan nilai ke-i dari masing-masing titik
- n yaitu jumlah dimensi data yang ada

4. Mengelompokkan data menurut jarak kedekatannya
5. Mencari nilai *Centeroid* yang baru

$$m_{i,j} = \frac{1}{n_i}\sum_{k=1}^{n_i} x_{k,j}$$

di mana:

- $m_{i,j}$ adalah *Centeroid* ke-i untuk variabel j
- n_i yaitu jumlah data *cluster* ke-i
- $x_{k,j}$ merupakan nilai data ke-k dalam *cluster* i untuk variabel ke-j

6. Ulangi langkah-langkah tersebut hingga tidak ada perubahan anggota dalam tiap cluster

Algoritma *K-Medoids*

K-Medoids merupakan algoritma clusterisasi partisional yang memiliki kemiripan dengan *K-Means*. Dalam kedua metode tersebut, pembagian data ke dalam k *cluster* dilakukan dengan tujuan meminimalkan jarak antara seluruh titik data dengan pusat cluster masing-masing. *K-Medoids* dikenal lebih tangguh dibandingkan *K-Means* saat diterapkan pada dataset berukuran besar. Perbedaan utama di antara keduanya dilihat pada proses penentuan pusat cluster. Pada *K-Means*, rata-rata koordinat dari titik-titik dalam suatu cluster digunakan sebagai pusat *cluster (centroid),* sedangkan pada *K-Medoids*, titik data aktual yang meminimalkan total jarak dipilih sebagai pusat *cluster (Medoids)*. Fungsi objektif dari *K-Medoids* dinyatakan secara matematis menggunakan jarak Euclidean kuadrat [13]. Proses algoritma *K-Medoids* sebagai berikut [14]:

1. Menetapkan jumlah *cluster* (k) yang akan dibentuk.
2. Menetapkan *Centeroid* awal secara random.
3. Menghitung jarak dari setiap data menggunakan metode *Euclidean Distance,* dengan rumus sebagai berikut:

$$D_{im} = \sqrt{\sum_{m}^{n}(x_{ij} - x_{mj})^2}$$

Metode pengukuran jarak *Euclidean* merupakan salah satu teknik yang umum diterapkan dalam analisis klaster. Karena metode ini cukup peka terhadap perbedaan skala antar variabel maupun keberadaan outlier, maka penting dilakukan proses normalisasi atau standarisasi data terlebih dahulu. Selain itu, outlier juga perlu ditangani dengan baik agar tidak menyebabkan hasil klaster menjadi tidak akurat.

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Dibandingkan dengan *K-Means,* algoritma *K-Medoids* cenderung lebih robust terhadap outlier karena pusat klasternya (medoid) berasal dari data aktual, bukan nilai rata-rata. Dengan demikian, pemilihan jenis jarak yang digunakan akan sangat memengaruhi kualitas pemisahan dan pembentukan klaster [15] (Sanmas et al., 2024).

4. Menentukan data dari setiap *cluster* sebagai kandidat *Medoids* baru.
5. Menghitung deviasi (S) dengan mengurangkan total jarak sebelumnya dengan total jarak setelah pemilihan *Medoids* baru. Jika nilai $S < 0$, maka objek tersebut digantikan dengan *Medoids* baru untuk membentuk himpunan *Medoids* yang diperbarui sebanyak k buah.
6. Langkah ini diulang hingga tidak terjadi perubahan *Medoids*.

Evaluasi

Salah satu metode yang sering dimanfaatkan untuk mengevaluasi kinerja algoritma klastering adalah *Davies-Bouldin Index* (DBI). Kualitas klaster yang terbentuk dinilai oleh DBI melalui perbandingan antara jarak antar klaster dan Tingkat penyebaran dalam masing-masing klaster. Ketika klaster yang dihasilkan bersifat semakin kompak dan memiliki jarak yang semakin jauh dari klaster lainnya, maka nilai DBI yang diperoleh akan semakin rendah. Nilai DBI yang kecil menunjukkan bahwa pemisahan antar klaster cukup jelas serta terdapat homogenitas yang tinggi dalam setiap klaster. Oleh karena itu, DBI digunakan sebagai indikator yang efisien dalam menilai sejauh mana struktur klaster yang dibentuk oleh suatu algoritma dapat dianggap optimal [16] :

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{i \neq j}(M_{i,j,\ldots,n})$$

HASIL DAN PEMBAHASAN

Statistik Deskriptif

Statistik deskriptif digunakan untuk merangkum informasi mengenai sebaran data berdasarkan masing-masing kategori pencemaran lingkungan, yang meliputi pencemaran air, tanah, udara, serta kategori tanpa pencemaran. Ringkasan hasil analisis tersebut disajikan dalam Tabel 2.

Tabel 2. Statistik Deskriptif.

Variable	Pencemaran Air	Pencemaran Tanah	Pencemaran Udara	Tidak Ada Pencemran
Minimum	7.000	1.000	4.000	157.000
Median	171.000	19.000	89.000	1126.000
Mean	290.000	27.610	131.680	1855.100
Maksimum	1366.000	122.000	583.000	7103.000
Standar Devaiasi	345.990	28.330	139.020	1806.320

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Pencemaran Air memiliki rata-rata sebesar 290 dengan standar deviasi yang cukup tinggi sebesar 345.990, menunjukkan adanya penyebaran data yang luas dengan nilai maksimum 1366. Sementara itu, Pencemaran Tanah menunjukkan distribusi yang lebih terkonsentrasi dengan rata-rata 27.610 dan standar deviasi 28.330, yang mencerminkan variasi yang relatif rendah antar provinsi. Pencemaran Udara memiliki rata-rata 131.680 dan standar deviasi 139.020, yang juga mengindikasikan adanya variasi yang signifikan antar wilayah. Adapun kategori "Tidak Ada Pencemaran" menunjukkan nilai rata-rata tertinggi, yaitu 1855.100, disertai standar deviasi paling besar sebesar 1806.320, yang menandakan bahwa distribusi data sangat tersebar dan berbeda jauh antar provinsi.

Uji Multikoliniearitas

Pengujian multikolinearitas dilakukan untuk mendeteksi adanya hubungan linear yang tinggi antar variabel independen dalam model regresi. Salah satu metode yang digunakan adalah perhitungan *Variance Inflation Factor* (VIF). Jika nilai VIF berada di bawah 10, maka dapat disimpulkan bahwa multikolinearitas tidak terjadi. Hasil perhitungan VIF disajikan pada Tabel 3.

Tabel 3. Nilai VIF.

Variable	Pencemaran Air	Pencemaran Tanah	Pencemaran Udara	Tidak Ada Pencemran
Nilai VIF	4.426	2.290	5.908	3.567

Berdasarkan hasil yang diperoleh, seluruh variabel independen menunjukkan nilai VIF di bawah 10. Dengan demikian, dapat disimpulkan bahwa tidak terdapat gejala multikolinearitas. Hal ini menunjukkan bahwa tidak ada hubungan linear yang kuat antar variabel bebas, sehingga model dapat dilanjutkan ke tahap analisis selanjutnya tanpa perlu perlakuan khusus terhadap multikolinearitas.

Penentuan Jumlah *Cluster*

Penentuan jumlah *cluster* yang optimal dilakukan dengan menggunakan metode *Davies-Bouldin Index* (DBI). Metode ini menilai kualitas *clustering* berdasarkan perbandingan antara jarak antar *cluster* dan penyebaran data di dalam masing-masing *cluster*. Angka DBI yang lebih rendah menunjukkan bahwa *cluster* yang terbentuk semakin kompak dan terpisah dengan jelas, sehingga hasil *clustering* dianggap lebih baik. Perhitungan DBI dilakukan untuk berbagai kemungkinan jumlah *cluster* pada algoritma *K-Means* dan *K-Medoids*. Hasil perhitungan tersebut ditampilkan dalam Tabel 4.

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Tabel 4. *Davies-Bouldin Index*.

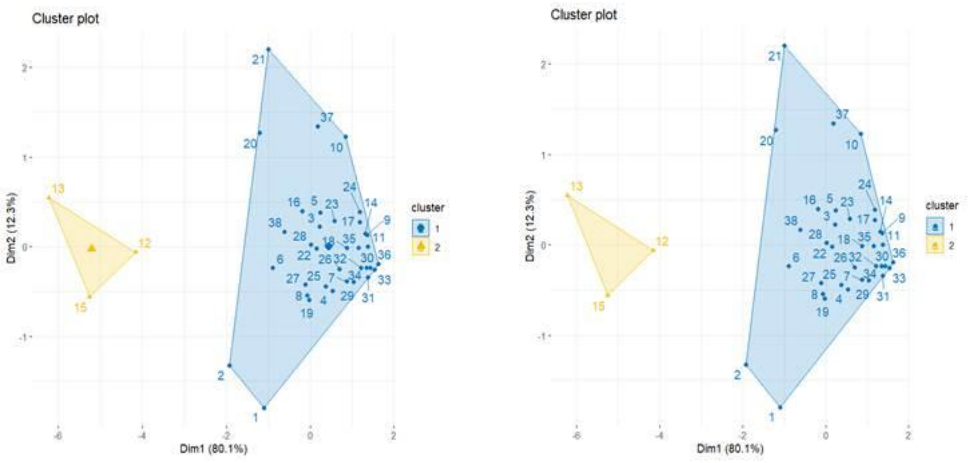
Metode	Jumlah Cluster	Davies-Bouldin Index
K-Means	**2**	**0.432**
	3	1.049
	4	0.929
	5	0.853
K-Medoids	**2**	**0.465**
	3	1.306
	4	1.066
	5	1.287

Berdasarkan hasil pada Tabel 4, nilai *Davies-Bouldin Index* terkecil untuk algoritma *K-Means* adalah sebesar 0.432 yang diperoleh saat jumlah *cluster* adalah 2. Hal ini menunjukkan bahwa pemodelan *clustering* terbaik menggunakan *K-Means* terjadi saat data dibagi menjadi 2 *cluster*, karena *cluster* yang terbentuk paling optimal. Demikian pula, pada algoritma *K-Medoids*, nilai DBI terkecil adalah 0.465, yang juga terjadi saat jumlah *cluster* adalah 2. Oleh karena itu, dapat disimpulkan bahwa jumlah *cluster* yang optimal untuk metode *K-Means* dan *K-Medoids* adalah 2 *cluster*. Pemilihan *K-Means* dan *K-Medoids* dalam penelitian ini didasarkan pada nilai *Davies-Bouldin Index* yang rendah, ini untuk menunjukkan bahwa pemisahan antar klaster cukup jelas serta terdapat homogenitas yang tinggi dalam setiap klaster, sehingga menjadikannya indikator yang efisien untuk menilai sejauh mana struktur klaster yang dibentuk dapat dianggap optimal (Wahidah et al., 2023).

Hasil Pengelompokan

Setelah dilakukan analisis penentuan *cluster* menggunakan metode *K-Means* dan *K-Medoids*, tahap selanjutnya adalah visualisasi hasil pengelompokan dari kedua algoritma yang dapat dilihat pada Gambar 2.

(a) (b)



Gambar 2. Visualsisasi *Cluster* (a) *K-Means* dan (b) *K-Medodis*

Dari Gambar 2a dan 2b dapat dilihat bahwa visualisasi dilakukan menggunakan jumlah *cluster* optimum, yaitu sebanyak dua *cluster*. Baik algoritma *K-Means* maupun *K-Medoids* menunjukkan hasil pengelompokan yang hampir sama,

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

dengan *Cluster* 1 beranggotakan 35 provinsi dan *Cluster* 2 terdiri atas 3 provinsi. Daftar lengkap dari masing-masing *cluster* disajikan pada Tabel 5 berikut.

Tabel 5. Hasil pengelompokkan *cluster* optimum

Cluster	Provinsi
Cluster 1	Aceh, Sumatera Utara, Sumatera Barat, Riau, Kep. Riau, Jambi, Bengkulu, Sumatera Selatan, Bangka Belitung, Lampung, DKI Jakarta, Banten, DIY, Bali, Nusa Tenggara Barat, NTT, Kalimantan Barat, Kalimantan Tengah, Kalimantan Selatan, Kalimantan Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Sulawesi Selatan, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku, Maluku Utara, Papua, Papua Barat, Papua Selatan, Papua Tengah, Papua Pegunungan, dan Papua Barat Daya.
Cluster 2	Jawa Tengah, Jawa Barat, dan Jawa Timur

Karaktersitik *Cluster*

Karakteristik masing-masing *cluster* dapat diinterpretasikan melalui nilai rata-rata variabel pencemaran lingkungan. Pada metode *K-Means* dan *K-Medoids*, karakteristik *cluster* diwakili oleh provinsi-provinsi yang paling menggambarkan *cluster* tersebut. Nilai yang ditampilkan merupakan rata-rata dari data jumlah desa/kelurahan yang telah distandarisasi menurut jenis pencemaran tertentu. Karakteristik ditampilkan pada Tabel 6 berikut.

Tabel 6. Karakteristik *cluster*.

Metode	*Cluster*	Pencemaran Air	Pencemaran Tanah	Pencemaran Udara	Tidak Ada Pencemaran
K-Means	1	-0.244	-0.192	-0.251	-0.203
	2	2.843	2.238	2.923	2.364
K-Medoids	1	-0.269	-0.092	-0.199	-0.584
	2	2.344	2.061	3.153	2.905

1. Metode *K-Means*
   - *Cluster* 1: *Cluster* ini menunjukkan nilai rata-rata negatif pada seluruh variabel pencemaran, yaitu pencemaran air (-0.244), pencemaran tanah (-0.192), pencemaran udara (-0.251), dan tidak ada pencemaran (-0.203). Nilai negatif ini merupakan hasil dari data yang telah distandarisasi. Hal ini mengindikasikan bahwa provinsi dalam *cluster* ini cenderung memiliki jumlah desa/kelurahan yang mengalami pencemaran relatif rendah dibanding rata-rata nasional.
   - *Cluster* 2: *Cluster* ini memiliki nilai rata-rata positif yang cukup tinggi, yakni pencemaran air (2.843), pencemaran tanah (2.238), pencemaran udara (2.923), dan tidak ada pencemaran (2.364). Ini menunjukkan bahwa provinsi dalam *cluster* ini cenderung memiliki jumlah desa/kelurahan yang mengalami pencemaran lebih tinggi dibandingkan rata-rata.
2. Metode *K-Medoids*

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

- *Cluster* 1: *Cluster* ini memiliki nilai rata-rata negatif pada seluruh variabel, khususnya pada variabel tidak ada pencemaran (-0.584), menunjukkan bahwa provinsi dalam *cluster* ini cenderung memiliki tingkat pencemaran lingkungan yang rendah atau jumlah desa/kelurahan yang mengalami pencemaran sedikit.
- *Cluster* 2: *Cluster* ini menunjukkan nilai rata-rata positif yang tinggi untuk semua jenis pencemaran, termasuk pencemaran udara (3.153), yang merupakan nilai tertinggi di antara semua variabel. Hal ini menunjukkan bahwa provinsi-provinsi dalam *cluster* ini memiliki jumlah desa/kelurahan yang tinggi pada semua jenis pencemaran

Dengan demikian, baik metode *K-Means* maupun *K-Medoids* sama-sama membentuk dua *cluster* utama berdasarkan jumlah desa/kelurahan yang mengalami jenis pencemaran tertentu yaitu, satu kelompok provinsi dengan tingkat pencemaran lingkungan yang rendah yang ditandai oleh nilai standar negatif, dan satu kelompok dengan tingkat pencemaran tinggi dengan nilai standar positif.

Provinsi yang memiliki banyak desa tercemar umumnya disebabkan oleh tingginya kepadatan penduduk, pesatnya urbanisasi, serta aktivitas industri dan pertanian intensif yang menghasilkan limbah mencemari air, tanah, dan udara. Ketiga provinsi dalam *Cluster* 2 Jawa Barat, Jawa Tengah, dan Jawa Timur merupakan pusat pertumbuhan ekonomi nasional dengan beban lingkungan yang tinggi akibat limbah domestik, industri, dan pertanian yang tidak terkelola dengan baik. Selain itu, konversi lahan, minimnya infrastruktur pengolahan limbah di desa-desa, lemahnya pengawasan lingkungan, serta rendahnya edukasi masyarakat terhadap pentingnya sanitasi dan pengelolaan sampah turut memperburuk kondisi pencemaran. Faktor-faktor ini menyebabkan desa-desa di provinsi tersebut lebih rentan terhadap pencemaran dibandingkan provinsi lain yang tergolong dalam *Cluster* 1, yang umumnya memiliki kepadatan penduduk lebih rendah, aktivitas industri yang terbatas, dan kondisi lingkungan yang masih relatif terjaga.

KESIMPULAN

Berdasarkan hasil analisis yang dilakukan, metode *K-Means* dan *K-Medoids* diuji dengan berbagai jumlah *cluster* (2 hingga 5), dan diperoleh bahwa konfigurasi terbaik terjadi pada saat menggunakan 2 cluster. Nilai *Davies-Bouldin Index* (DBI) terkecil dicapai oleh metode *K-Means* sebesar 0.432, menjadikannya algoritma yang lebih optimal dibanding *K-Medoids* dengan DBI sebesar 0.465. Menariknya, kedua metode menghasilkan pembagian cluster yang sama, yaitu *Cluster* 1 yang terdiri atas 35 provinsi dengan tingkat pencemaran lingkungan relatif rendah, dan *Cluster* 2 yang hanya berisi 3 provinsi, yakni Jawa Barat, Jawa Tengah, dan Jawa Timur, yang menunjukkan tingkat pencemaran lebih tinggi. Tingginya tingkat pencemaran di provinsi-provinsi *Cluster* 2 disebabkan oleh berbagai faktor, antara lain kepadatan penduduk yang sangat tinggi, pesatnya urbanisasi, dominasi aktivitas industri dan pertanian intensif, lemahnya pengawasan lingkungan, konversi lahan tanpa tata kelola yang baik, serta kurangnya infrastruktur pengolahan limbah dan edukasi masyarakat terkait lingkungan. Kondisi ini menyebabkan desa-desa di provinsi tersebut lebih

rentan terhadap pencemaran, sehingga memerlukan perhatian khusus dan peningkatan upaya pengendalian pencemaran secara terpadu dan berkelanjutan.

DAFTAR PUSTAKA

[1] M. Anjelita, A. P. Windarto, A. Wanto, And I. Sudahri, "Pengembangan Datamining Klastering Pada Kasus Pencemaran Lingkungan Hidup," Seminar Nasional Teknologi Komputer & Sains (Sainteks), 2020, Pp. 309–313.

[2] I. Mardianti, K. Kasmantoni, And A. Walid, "Pengembangan Modul Pembelajaran Ipa Berbasis Etnosains Materi Pencemaran Lingkungan Untuk Melatih Literasi Sains Siswa Kelas Vii Di Smp," *Bio-Edu: Jurnal Pendidikan Biologi*, Vol. 5, No. 2, Pp. 98–107, Aug. 2020, Doi: 10.32938/Jbe.V5i2.545.

[3] R. Maulana, "Analisis Tingkat Pencemaran Lingkungan Pada Kota/Kabupaten Di Jawa Tengah Menggunakan Metode K-Means," *Jurnal Informatika Dan Teknik Elektro Terapan*, Vol. 11, No. 3, Aug. 2023, Doi: 10.23960/Jitet.V11i3.3177.

[4] N. Rohman And A. Wibowo, "Perbandingan Metode K-Medoids Dan Metode K-Means Dalam Analisis Segmentasi Pelanggan Mall," *Sintech Journal*, Vol. 7, No. 1, Pp. 49–58, 2024, Doi: Https://Doi.Org/10.31598.

[5] M. Bhakti, J. Dedy Irawan, And D. Rudhistiar, "Pengelompokan Game Berdasarkan Data Top Seller Pada Website 'Steam' Menggunakan Metode K-Medoids," *Jurnal Mahasiswa Teknik Informatika*, Vol. 8, No. 5, Pp. 8242–8249, 2024, [Online]. Available: Https://Store.Steampowered.Com/

[6] A. T. Sipayung, Saifullah, And W. Riki, "Penerapan Metode K-Means Dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Jenis Pencemaran Lingkungan Hidup Berdasarkan Provinsi," *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, Vol. 1, No. 4, Pp. 104–111, 2020, [Online]. Available: Http://Www.Bps.Go.Id.

[7] G. Mardiatmoko, "The Importance Of The Classical Assumption Test In Multiple Linear Regression Analysis (A Case Study Of The Preparation Of The Allometric Equation Of Young Walnuts)," *Barekeng*, Vol. 14, No. 3, Pp. 333–342, Sep. 2020, Doi: 10.30598/Barekengvol14iss3pp333-342.

[8] I. N. Azizah, P. R. Arum, And R. Wasono, "Model Terbaik Uji Multikolinearitas Untuk Analisis Faktor-Faktor Yang Mempengaruhi Produksi Padi Di Kabupaten Blora Tahun 2020," Prosiding Seminar Nasional Unimus, 2021, Pp. 61–69.

[9] D. S. Purba, W. J. Tarigan, M. Sinaga, And V. Tarigan, "Pelatihan Penggunaan Software Spss Dalam Pengolahan Regressi Linear Berganda Untuk Mahasiswa Fakultas Ekonomi Universitas Simalungun Di Masa Pandemi Covid 19," *Jurnal Karya Abdi*, Vol. 5, No. 2, Pp. 202–208, Aug. 2021.

[10] D. Ramdhan, G. Dwilestari, R. Danar Dana, And A. Ajiz, "Clustering Data Persediaan Barang Dengan Menggunakan Metode K-Means.," *Means (Media Informasi Analisa Dan Sistem)*, Vol. 7, No. 1, Pp. 1–9, 2022, Doi: Http://Ejournal.Ust.Ac.Id/Index.Php/Jurnal_Means/.

[11] A. Fira, C. Rozikin, And Garno, "Komparasi Algoritma K-Means Dan K-Medoids Untuk Pengelompokkan Penyebaran Covid-19 Di Indonesia," *Journal Of Applied Informatics And Computing (Jaic)*, Vol. 5, No. 2, Pp. 133–138, Dec. 2021, Doi: Http://Jurnal.Polibatam.Ac.Id/Index.Php/Jaic.

[12] A. Sulistiyawati And E. Supriyanto, "Implementasi Algoritma K-Means Clustring Dalam Penetuan Siswa Kelas Unggulan," *Jurnal Tekno Kompak*, Vol. 15, No. 2, 2020.

[13] A. Sobrinho Campolina Martins, L. Ramos De Araujo, And D. Rosana Ribeiro Penido, "K-Medoids Clustering Applications For High-Dimensionality Multiphase Probabilistic

Prosiding Seminar Nasional UNIMUS
Volume 8, 23 Oktober 2025
e-ISSN: 2654-3168 | p-ISSN: 2654-3257

Power Flow," *International Journal Of Electrical Power And Energy Systems*, Vol. 157, Jun. 2024, Doi: 10.1016/J.Ijepes.2024.109861.

[14] N. Wahidah, O. Juwita, And F. Nurman Arifin, "Pengelompokkan Daerah Rawan Bencana Di Kabupaten Jember Menggunakan Metode K-Means Clustering," 2023.

[15] S. A. Sanmas, R. Nurmalita, D. Sulistiyani, And M. Al Haris, "Pengelompokkan Wilayah Banjir Di Jawa Tengah Untuk Mitigasi Banjir Menggunakan Pendekatan K-Medoids," *Jurnal Statistika Dan Komputasi*, Vol. 3, No. 2, Pp. 51–61, Dec. 2024, Doi: 10.32665/Statkom.V3i2.3223.

[16] Sekar Setyaningtyas, B. Indarmawan Nugroho, And Z. Arif, "Tinjauan Pustaka Sistem Pada Data Mining: Studi Kasus Algotitma K-Meansclustering," *Jurnal Teknoif Teknik Informatika Institut Teknologi Padang*, Vol. 10, No. 2, Pp. 52–61, Oct. 2022, Doi: 10.21063/Jtif.2022.V10.2.52-61.