

Pemetaan Risiko Penyakit Diare di Jawa Timur dengan Analisis Kluster K-Means Berbasis Indikator Sosial dan Kesehatan

Diana Novitasari¹, Amellia Harmaimun Hidayah², Dwi Arman Prasetya³

^{1,3}Diana Novitasari, Dwi Arman Prasetya (Sains Data, Universitas Pembangunan Nasional “Veteran” Jawa Timur)

¹23083010014@student.upnjatim.ac.id

³arman.prasetya.sada@upnjatim.ac.id

²Amellia Harmaimun Hidayah (Sains Data, Universitas Pembangunan Nasional “Veteran” Jawa Timur)

Corresponding author email: 23083010034@student.upnjatim.ac.id

ABSTRAK

Penyakit diare masih menjadi salah satu permasalahan kesehatan masyarakat utama di Indonesia, terutama di Provinsi Jawa Timur yang mencatat jumlah kasus tinggi dengan distribusi tidak merata antarwilayah. Ketimpangan akses air bersih, sanitasi, dan fasilitas kesehatan berperan besar dalam perbedaan tingkat risiko penyakit ini. Penelitian ini bertujuan untuk mengelompokkan kabupaten/kota di Jawa Timur berdasarkan tingkat risiko penyakit diare menggunakan metode *K-Means Clustering*. Penelitian bersifat kuantitatif deskriptif dengan data sekunder dari Badan Pusat Statistik tahun 2023 yang mencakup 38 kabupaten/kota. Variabel yang digunakan meliputi jumlah kasus diare, akses air minum dan sanitasi layak, tingkat kemiskinan, kepadatan penduduk, dan jumlah puskesmas. Analisis dilakukan melalui tahapan pra-pemrosesan data, standarisasi dengan metode *Z-Score*, penentuan jumlah kluster optimal menggunakan metode *Elbow* dan *Silhouette Score*, serta evaluasi dengan *Calinski Harabasz* dan *Davies Bouldin Index*. Hasil penelitian menghasilkan lima kluster dengan tingkat risiko berbeda. Wilayah Tapal Kuda dan Madura termasuk kategori risiko tinggi, sedangkan kota-kota besar seperti Malang dan Mojokerto tergolong rendah. Temuan ini menegaskan bahwa *K-Means* mampu menggambarkan pola spasial risiko diare secara komprehensif dan dapat menjadi acuan dalam penyusunan kebijakan kesehatan berbasis data dan pemerataan pelayanan publik di Jawa Timur.

Keywords: Diare, *K-Means Clustering*, Sanitasi, Sosial Ekonomi, Jawa Timur

I. PENDAHULUAN

Penyakit diare merupakan salah satu masalah kesehatan masyarakat yang masih menjadi tantangan serius di Indonesia. Penyakit ini ditandai dengan meningkatnya frekuensi buang air besar lebih dari tiga kali per hari dengan konsistensi lembek atau cair, serta berisiko menimbulkan dehidrasi berat hingga kematian, terutama pada anak-anak dan balita. Berdasarkan Profil Kesehatan Indonesia 2022, penyakit diare termasuk dalam kelompok dengan angka kejadian tinggi dan berkontribusi besar terhadap beban kesakitan nasional. Meskipun dapat dicegah, tingginya kasus di berbagai wilayah menunjukkan adanya ketimpangan kondisi kesehatan masyarakat.

Diare tidak hanya berkaitan dengan aspek medis, tetapi juga mencerminkan kondisi sosial dan lingkungan suatu daerah. Faktor yang memengaruhi di antaranya adalah kualitas air minum, sanitasi lingkungan, perilaku hidup bersih, kepadatan penduduk, kemiskinan, dan ketersediaan fasilitas kesehatan. Dalam penelitian sebelumnya telah menunjukkan bahwa rendahnya akses air bersih dan sanitasi layak berhubungan signifikan dengan meningkatnya kasus diare pada balita di Indonesia [1]. Sementara itu, Badan Pusat Statistik (2023) melaporkan bahwa akses air minum layak nasional mencapai 92,64%, sedangkan sanitasi layak 83,60%, menandakan masih

adanya kesenjangan pelayanan dasar di sejumlah daerah. Kondisi ini memperkuat pandangan bahwa ketersediaan infrastruktur dasar berperan penting dalam pencegahan penyakit menular berbasis lingkungan.

Provinsi Jawa Timur merupakan salah satu wilayah dengan beban kasus diare yang tinggi di Indonesia. Berdasarkan data BPS Provinsi Jawa Timur (2023) yang bersumber dari Dinas Kesehatan Jawa Timur, pada tahun 2022 tercatat ribuan kasus diare di setiap kabupaten/kota, misalnya Kota Surabaya mencapai 22.327 kasus. Variasi ini menunjukkan bahwa penyebaran penyakit tidak merata dan cenderung dipengaruhi oleh perbedaan kondisi sosial, ekonomi, serta ketersediaan layanan kesehatan antarwilayah. Wilayah perkotaan seperti Surabaya dan Sidoarjo memiliki fasilitas kesehatan yang memadai, sedangkan kabupaten masih menghadapi keterbatasan akses air bersih, sanitasi, dan layanan kesehatan dasar.

Dalam penelitian kesehatan masyarakat, analisis kluster menjadi pendekatan efektif untuk memahami pola sebaran penyakit berbasis wilayah. Metode ini dapat mengelompokkan daerah dengan karakteristik serupa sehingga kebijakan intervensi dapat dilakukan lebih tepat sasaran. Salah satu algoritma yang banyak digunakan adalah *K-Means Clustering*, karena efisien dan mudah diinterpretasikan. Penelitian sebelumnya di Provinsi Jawa Barat membuktikan bahwa efektivitas metode ini dalam memetakan tingkat risiko diare berdasarkan indikator sosial dan lingkungan [2]. Namun, kajian serupa untuk konteks Provinsi Jawa Timur masih jarang dilakukan, padahal wilayah ini memiliki keragaman sosial dan demografi yang tinggi. Oleh karena itu, penelitian ini bertujuan untuk menganalisis dan memetakan wilayah berisiko penyakit diare di Provinsi Jawa Timur menggunakan metode *K-Means Clustering* dengan mempertimbangkan indikator sosial dan kesehatan, yaitu jumlah kasus diare, persentase rumah tangga dengan akses air minum layak, sanitasi layak, tingkat kemiskinan, kepadatan penduduk, dan jumlah puskesmas. Hasil penelitian diharapkan dapat menjadi dasar penyusunan kebijakan kesehatan berbasis bukti serta mendukung pemerataan pelayanan kesehatan masyarakat di Jawa Timur secara berkeadilan.

II. METODOLOGI PENELITIAN

Bab ini menguraikan secara rinci langkah-langkah yang ditempuh dalam pelaksanaan penelitian. Pembahasannya mencakup jenis dan pendekatan penelitian, sumber serta jenis data yang digunakan, hingga metode analisis yang diterapkan. Uraian metodologi ini dimaksudkan untuk memberikan gambaran yang utuh mengenai proses penelitian, sehingga hasil yang diperoleh dapat dipahami dan dipertanggungjawabkan secara ilmiah.

2.1. Jenis dan Pendekatan Penelitian

Penelitian ini merupakan penelitian kuantitatif deskriptif dengan pendekatan data mining menggunakan algoritma *K-Means Clustering*. Tujuan penelitian ini adalah mengelompokkan 38 kabupaten/kota di Provinsi Jawa Timur berdasarkan indikator sosial dan kesehatan yang berpengaruh terhadap risiko penyakit diare. Pemilihan metode *K-Means* didasarkan pada kemampuannya mengelompokkan data numerik secara efisien dan menghasilkan kluster yang mudah diinterpretasikan untuk mendukung analisis spasial penyakit berbasis wilayah.

2.2. Jenis dan Sumber Data

Penelitian ini menggunakan data sekunder kuantitatif yang diperoleh dari Badan Pusat Statistik (BPS) tahun 2023. Data mencakup 38 kabupaten dan kota dengan unit analisis wilayah administratif. Seluruh data bersumber dari publikasi resmi yang tersedia melalui situs bps.go.id. Data tersebut diolah menggunakan metode *K-Means Clustering* untuk mengidentifikasi pola wilayah berdasarkan kesamaan karakteristik sosial dan

kesehatan yang memengaruhi tingkat risiko penyakit diare. Tabel berikut menyajikan variabel penelitian yang digunakan dalam analisis:

Tabel 1. Variabel Penelitian

Variabel	Keterangan
x_1	Jumlah Kasus Diare
x_2	Persentase Rumah Tangga dengan Akses Air Minum Layak
x_3	Persentase Rumah Tangga dengan Akses Sanitasi Layak
x_4	Persentase Penduduk Miskin
x_5	Angka Kepadatan Penduduk
x_6	Jumlah Puskesmas

Seluruh variabel tersebut dipilih karena memiliki relevansi kuat terhadap penyebaran penyakit diare. Akses air dan sanitasi mencerminkan kondisi lingkungan, sedangkan tingkat kemiskinan dan kepadatan penduduk menggambarkan kondisi sosial ekonomi yang turut memengaruhi risiko penularan. Sementara itu, jumlah puskesmas dan jumlah kasus diare digunakan sebagai indikator utama untuk menilai kondisi kesehatan masyarakat di setiap kabupaten/kota.

2.3. Metode Analisis Data

Metode analisis data dalam penelitian ini terdiri atas beberapa tahapan yang saling berkaitan, dimulai dari pra-pemrosesan data (data *preprocessing*), standarisasi data, penentuan jumlah kluster optimal, penerapan algoritma *K-Means*, hingga evaluasi dan visualisasi hasil klusterisasi. Setiap tahapan dirancang untuk memastikan keakuratan dan reliabilitas hasil analisis yang diperoleh.

2.3.1. Data Preprocessing

Tahap pra-pemrosesan data dilakukan untuk memastikan bahwa data yang digunakan layak dan siap untuk dianalisis. Proses ini meliputi beberapa langkah utama, yaitu pemeriksaan kelengkapan data (*missing values*), pembersihan format kolom, serta penyesuaian tipe data agar seluruh variabel dapat diolah secara numerik. Selain itu, dilakukan juga peninjauan terhadap kesesuaian rentang nilai dan satuan antarvariabel agar tidak terjadi ketidakkonsistenan dalam pengolahan. Seluruh tahapan pra-pemrosesan dirancang untuk meminimalkan kesalahan teknis pada tahap analisis berikutnya. Proses ini dilakukan menggunakan perangkat lunak *Python*, yang memungkinkan pengelolaan data secara efisien dan akurat sebelum dilakukan standarisasi dan pengelompokan kluster.

2.3.2. Standarisasi Data

Standarisasi data dilakukan untuk menyamakan skala dan satuan antarvariabel agar dapat dibandingkan secara proporsional dalam proses analisis [1]. Tahapan ini penting karena perbedaan skala pada setiap variabel dapat menyebabkan hasil analisis menjadi bias, terutama pada metode yang sensitif terhadap jarak, seperti *K-Means Clustering*. Penelitian ini menggunakan metode *Z-Score normalization*, yaitu teknik yang mengubah setiap nilai variabel menjadi satuan baku berdasarkan rata-rata dan simpangan baku variabel tersebut. Proses ini dinyatakan dalam rumus [1]:

$$z = \frac{\chi_i - \mu}{\sigma} \quad (1)$$

Keterangan:

z = nilai *Z-Score*

χ = nilai observasi

μ = rata-rata

σ = standar deviasi

2.3.3. Penentuan Jumlah Kluster Optimal

Penentuan jumlah kluster optimal merupakan tahap penting untuk memperoleh hasil pengelompokan yang representatif dan tidak bias. Dalam algoritma *K-Means*, jumlah kluster (k) harus ditentukan sebelum proses iterasi, sehingga pemilihan nilai yang tepat sangat berpengaruh terhadap kualitas hasil akhir. Dalam penelitian ini, dua pendekatan digunakan untuk menentukan jumlah kluster terbaik, yaitu *Elbow Method* dan *Silhouette Coefficient* [3].

- a. Pertama, metode *Elbow* dilakukan dengan mengamati titik “siku” (*elbow point*) pada grafik hubungan antara jumlah kluster dan nilai *Within-Cluster Sum of Squares* (WCSS). Titik siku menunjukkan bahwa penambahan kluster selanjutnya tidak lagi memberikan pengurangan WCSS yang signifikan. Rumus WCSS dinyatakan sebagai berikut [4]:

$$SSE = \sum_{k=1}^K \sum_{x_i \in S_k} \|x_i - C_k\|^2 \quad (2)$$

Keterangan:

x_i : data ke- i yang termasuk dalam kluster ke- k

S_k : himpunan objek atau anggota yang berada pada kluster ke- k

C_k : pusat kluster (*centroid*) ke- k

- b. Kedua, *Silhouette Coefficient* mengukur seberapa baik suatu objek terkelompok dibandingkan dengan kluster lain. Nilainya dihitung dengan rumus [5]:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (3)$$

Keterangan:

$a(i)$ = rata-rata jarak antara data ke- i dan semua data lain dalam kluster yang sama

$b(i)$ = rata-rata jarak antara data ke- i dan semua data lain dalam kluster terdekat berikutnya

Nilai k yang menghasilkan rata-rata *Silhouette Coefficient* tertinggi dianggap sejumlah kluster optimal. Hasil dari kedua metode tersebut digunakan secara bersamaan untuk menentukan jumlah kluster yang paling stabil dan bermakna dalam konteks data sosial dan kesehatan di Provinsi Jawa Timur.

2.3.4. Proses Klasterisasi Menggunakan *K-Means*

Metode *K-Means Clustering* merupakan salah satu algoritma pengelompokan data non-hirarki yang digunakan untuk membagi sekumpulan objek ke dalam beberapa kluster berdasarkan tingkat kemiripan karakteristiknya [5].

1. Menentukan Jumlah Kluster (k)

Tahap awal adalah menentukan jumlah kluster (k) yang akan dibentuk. Nilai k ditentukan secara empiris menggunakan metode *Elbow* dan *Silhouette Coefficient* untuk memperoleh jumlah kluster yang paling sesuai dengan karakteristik data. Pemilihan nilai k ini sangat penting karena memengaruhi representasi pola distribusi risiko penyakit diare di Provinsi Jawa Timur.

2. Inisialisasi *Centroid* Awal

Setelah jumlah kluster ditentukan, algoritma akan memilih k titik awal secara acak sebagai pusat kluster sementara (initial *centroid*). Titik ini berfungsi sebagai acuan awal untuk proses perhitungan jarak antar data.

3. Menghitung Jarak ke *Centroid*

Setiap data pada dataset dihitung jaraknya terhadap seluruh *centroid* menggunakan metrik *Euclidean Distance*. Rumusnya adalah sebagai berikut [2]:

$$dEuclidean = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (4)$$

Keterangan

a_i = nilai ke-i dari objek pertama

b_i = nilai ke-i dari objek kedua

n = jumlah variabel data yang dibandingkan

4. Pengelompokan Objek Berdasarkan *Centroid* Terdekat

Setiap data dialokasikan ke dalam kluster dengan jarak paling dekat terhadap *centroid* hasil perhitungan sebelumnya. Proses ini memastikan bahwa objek-objek dengan karakteristik serupa tergabung dalam kluster yang sama.

5. Pembaruan *Centroid*

Setelah seluruh data terkelompok, *centroid* baru dihitung kembali dengan mengambil rata-rata posisi seluruh data dalam kluster yang sama. Proses ini bertujuan untuk menyesuaikan pusat kluster agar semakin merepresentasikan anggota-anggotanya [6].

$$C_k = \frac{1}{n_k} \sum d_i \quad (5)$$

Keterangan:

C_k = posisi *centroid* baru untuk kluster ke-k

n_k = jumlah data atau anggota dalam kluster-k

d_i = nilai data ke-i yang tergabung dalam kluster ke-k

6. Iterasi hingga *Konvergen*

Langkah perhitungan jarak dan pembaruan *centroid* diulang hingga posisi *centroid* tidak mengalami perubahan signifikan atau telah mencapai batas iterasi maksimum. Kondisi ini disebut konvergensi, yang menandakan bahwa proses klusterisasi telah stabil dan hasil pengelompokan akhir dapat digunakan untuk analisis lebih lanjut.

Proses di atas dilakukan menggunakan bahasa pemrograman *Python* dengan pustaka *Scikit-Learn*, yang menyediakan fungsi *KMeans()* untuk mempermudah penerapan algoritma secara efisien.

2.3.5. Evaluasi Hasil Klusterisasi

Tahap evaluasi hasil klusterisasi dilakukan untuk menilai kualitas dan ketepatan hasil pengelompokan setelah jumlah kluster optimal diperoleh. Dua metrik yang digunakan dalam penelitian ini adalah *Calinski-Harabasz Index* (CH) dan *Davies-Bouldin Index* (DBI).

- a. *Calinski-Harabasz Index* (CH) mengevaluasi rasio variansi antar kluster terhadap variansi dalam kluster [7]:

$$CH = \frac{(SSB)}{(SSW)} \times \frac{N - k}{k - 1} \quad (6)$$

$$SSW = \sum_{i=k}^k \sum_{x_i \in C_i} (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \quad (7)$$

$$SSB = \sum_{i=k}^k N_i (x_i - \bar{x}_i)(x_i - \bar{x}_i)^T \quad (8)$$

Keterangan:

CH : nilai indeks *Calinski-Harabasz*
 SSB : *sum of square between cluster* (variasi antar kluster)
 SSW : *sum of square within cluster* (variasi dalam kluster)
 N_i : jumlah data dalam kluster- i
 k : jumlah kluster
 x_i : vektor data ke- i dalam kluster
 $\bar{x}_i$: *centroid* (rata-rata) dari kluster ke- i
 C_i : himpunan data pada kluster ke- i
 $(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T$: jarak kuadrat antara data dengan *centroid*
 Nilai CH yang tinggi menandakan pemisahan antar kluster yang baik.

- b. *Davies-Bouldin Index* (DBI) digunakan untuk mengukur kemiripan antar kluster dengan rumus [8]:

$$DBI = \frac{1}{N} \sum_{i=1}^N R_i \quad (9)$$

Keterangan:

DBI : indeks yang merepresentasikan tingkat kualitas hasil proses klusterisasi; semakin kecil nilainya, semakin baik tingkat pemisahan antar kluster
 N : jumlah keseluruhan kluster yang terbentuk
 R_i : rasio kemiripan antara kluster ke- i dengan kluster lain yang paling mendekati, yang dirumuskan sebagai:

$$R_i = \max_{j \neq i} \frac{S_i + S_j}{M_{i,j}} \quad (10)$$

Keterangan:

S_i : nilai rerata jarak antara setiap objek dalam kluster ke- i terhadap pusat klasternya, yang menunjukkan tingkat kekompakan kluster
 S_j : nilai rata-rata jarak antara setiap objek dalam kluster ke- j terhadap pusat klasternya
 $M_{i,j}$: jarak antara pusat kluster ke- i dan kluster ke- j , yang menggambarkan tingkat keterpisahan antar kluster
 Semakin kecil nilai DBI , semakin baik kualitas pemisahan antar kluster.

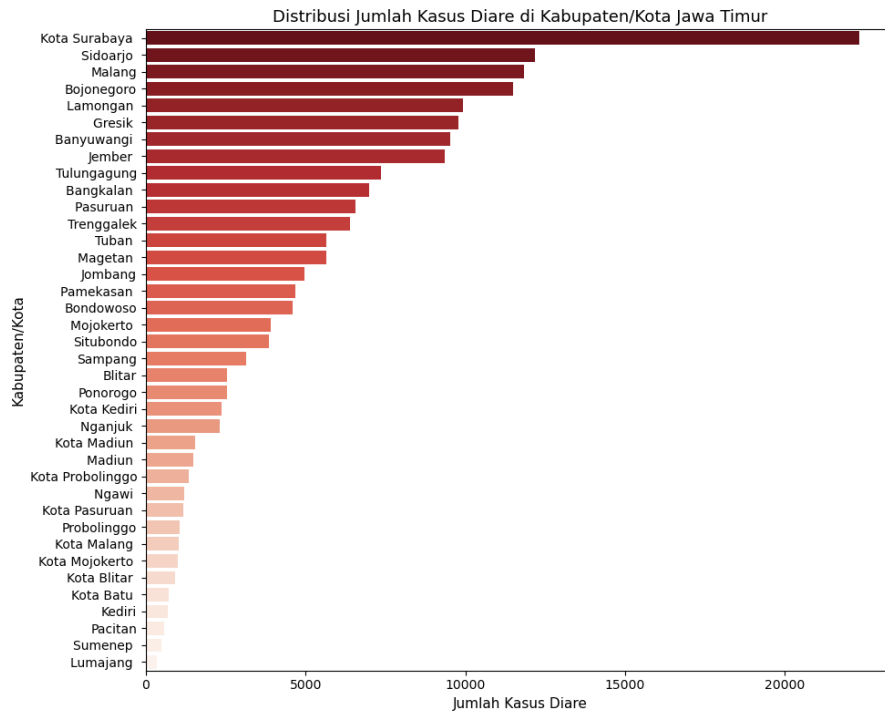
2.3.6. Visualisasi Hasil Klusterisasi

Visualisasi hasil klusterisasi dilakukan untuk menampilkan pola sebaran dan karakteristik tiap kluster yang terbentuk dari algoritma *K-Means*. Teknik visualisasi yang digunakan meliputi *Principal Component Analysis* (PCA) dua dan tiga dimensi untuk mereduksi dimensi data serta menampilkan distribusi antar kluster secara lebih sederhana. Selain itu, diagram batang dan peta sebaran spasial digunakan untuk memperlihatkan perbandingan rata-rata setiap indikator sosial dan kesehatan antar kluster. Seluruh visualisasi dihasilkan menggunakan bahasa pemrograman *Python* dengan pustaka *Matplotlib*, *Seaborn*, dan *Geopandas* guna memperkuat interpretasi hasil pengelompokan wilayah berdasarkan tingkat risiko penyakit diare di Provinsi Jawa Timur.

III. HASIL DAN ANALISIS

3.1. Pendistribusian Jumlah Kasus Diare

Provinsi Jawa Timur meliputi 38 kabupaten/kota. Berdasarkan visualisasi pada Gambar 1, penyebaran total kasus diare di setiap wilayah administratif menunjukkan variasi yang jelas. Lima daerah dengan kasus diare tertinggi di semua kelompok usia meliputi Kota Surabaya, Sidoarjo, Malang, Bojonegoro, dan Lamongan.



Gambar 1. Persebaran Kasus Diare di Provinsi Jawa Timur Tahun 2023

Sebelum melakukan analisis klusterisasi, statistik deskriptif yang disajikan pada Tabel 1 telah diverifikasi untuk memastikan kelengkapan data. Semua variabel yang digunakan dalam penelitian ini terbukti lengkap (tanpa adanya *missing data*), meliputi Jumlah Kasus Diare (x_1), Persentase Rumah Tangga dengan Akses Air Minum Layak (x_2), Persentase Rumah Tangga dengan Akses Sanitasi Layak (x_3), Persentase Penduduk Miskin (x_4), Angka Kepadatan Penduduk (x_5), dan Jumlah Puskesmas (x_6).

3.2. Statistika Deskriptif

Analisis Statistika Deskriptif ini melibatkan enam variabel indikator sosial dan kesehatan di 38 kabupaten/kota di Provinsi Jawa Timur. Parameter yang digunakan meliputi nilai minimum, maksimum, rerata, dan simpangan baku. Hasil dari analisis statistik deskriptif tersebut tersaji secara rinci dalam tabel dibawah ini.

Tabel 2. Statistik Deskriptif Secara Umum

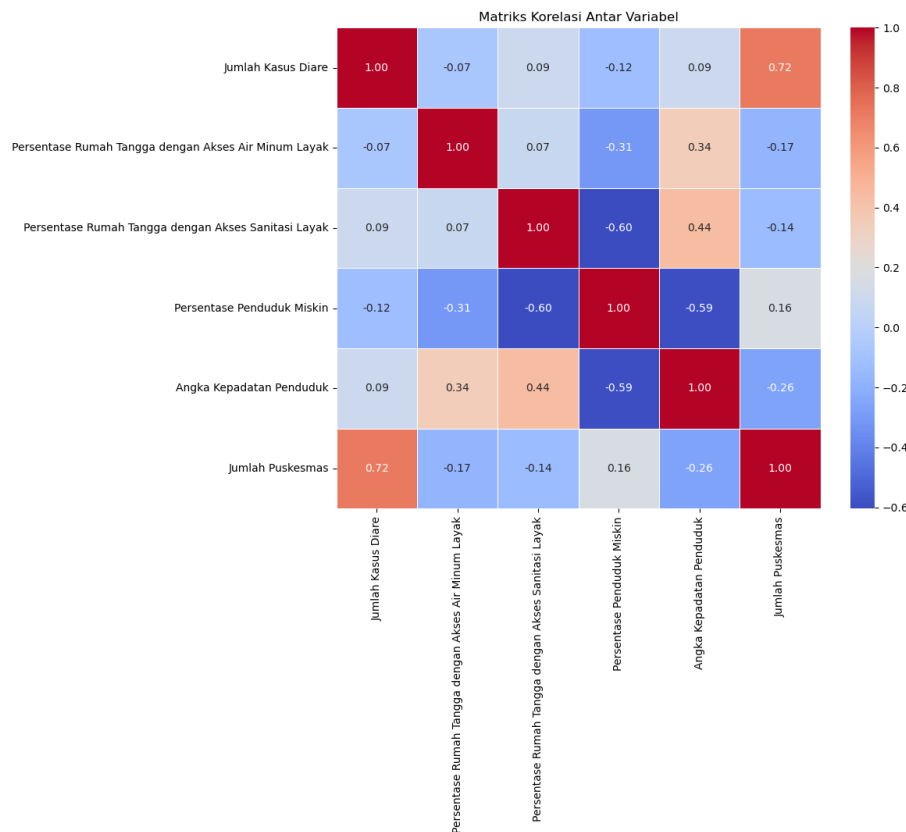
Variabel	count	mean	std	min	max	25%	50%	75%
x_1	38	4824.68	4659.22	328.0	22327.0	1179.0	3484.5	6873.5
x_2	38	96.12	4.81	79.26	100.0	95.69	97.61	98.84
x_3	38	84.59	12.18	50.3	98.18	81.59	87.84	93.77
x_4	38	10.33	4.27	3.79	21.61	7.26	9.63	12.45

x_5	38	1937.92	2262.03	410.0	8667.0	644.5	855.5	1204.25
x_6	38	25.58	12.85	3.0	63.0	20.25	25.0	32.75

Angka yang tersaji dalam Tabel 2 diatas menunjukkan bahwa rerata jumlah kasus diare sejumlah 4.825 kasus. Rerata persentase rumah tangga dengan akses air minum layak sebesar 96,12%. Rerata persentase rumah tangga dengan akses sanitasi layak berkisar 84,59%. Selain itu, untuk rerata persentase penduduk miskin sebesar 10,33%. Angka kepadatan penduduk menunjukkan rerata sebesar 1.938 ribu jiwa per km^2 , sementara itu jumlah puskesmas menghasilkan rerata sebesar 26 unit di setiap kabupaten/kota.

3.3. Korelasi Antar Variabel

Pengujian korelasi diterapkan hanya pada variabel-variabel yang relevan untuk mengukur derajat dan arah hubungan di antara setiap indikator. Hasil pengukuran koefisien ini, yang merangkum interaksi antar variabel, kemudian divisualisasikan dalam *Heatmap* Matriks Korelasi.



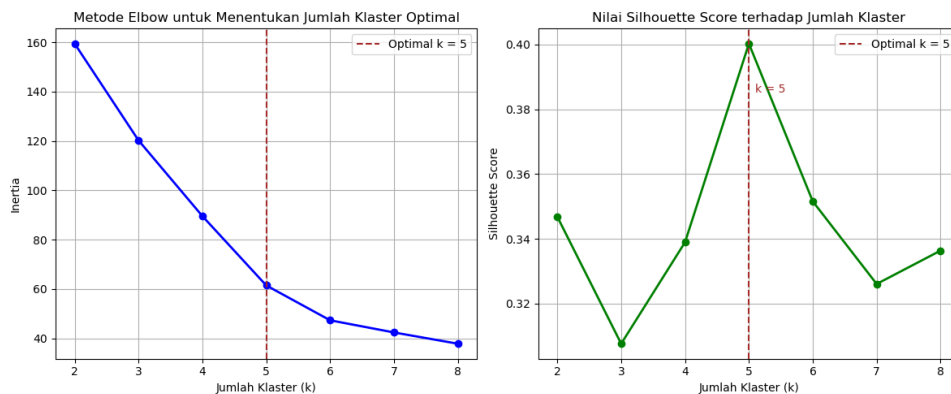
Gambar 2. Visualisasi Korelasi Antar Variabel

Hasil analisis korelasi pada Gambar 2 menunjukkan variasi hubungan antar indikator sosial dan kesehatan di Jawa Timur. Korelasi positif kuat terjadi antara jumlah kasus diare dan jumlah puskesmas ($r = 0,72$) menunjukkan bahwa daerah dengan fasilitas kesehatan lebih banyak cenderung memiliki laporan kasus yang lebih lengkap, mencerminkan kapasitas deteksi yang lebih baik. Teridentifikasi korelasi negatif yang kuat ($r = -0,60$) antara tingginya angka kemiskinan dan rendahnya persentase akses sanitasi layak. Temuan ini secara nyata menunjukkan bahwa keterbatasan ekonomi pada tingkat rumah tangga berdampak langsung sebagai hambatan utama dalam peningkatan kualitas sanitasi. Selanjutnya, meskipun terindikasi bahwa daerah perkotaan (urban) memiliki keunggulan infrastruktur dasar, yang ditunjukkan oleh

korelasi positif sedang antara kepadatan penduduk dengan akses air minum ($r = 0,34$) dan sanitasi layak ($r = 0,44$), adanya populasi yang padat tidak dapat diabaikan sebagai faktor risiko penyebaran penyakit yang tetap saling terkait.

3.4. Penentuan Cluster Optimal

Setelah menganalisis hubungan antarvariabel, tahap berikutnya adalah menentukan jumlah kluster yang paling representatif untuk menggambarkan karakteristik wilayah berdasarkan indikator sosial dan kesehatan. Penentuan jumlah kluster optimal dilakukan menggunakan pendekatan *Metode Elbow* dan *Silhouette Score* sebagaimana ditampilkan pada Gambar 3.



Gambar 3. Grafik k Optimal antara Metode *Elbow* dengan Nilai *Silhouette Score*

Jumlah kluster optimal ditentukan melalui analisis komparatif antara Metode *Elbow* dan *Silhouette Score*. Gambar 3 menunjukkan bahwa titik belok (*elbow*) terlihat jelas pada $k=5$, yang mengisyaratkan penurunan inersia mulai melambat. Penentuan ini divalidasi oleh nilai *Silhouette Score* tertinggi (0,400) pada $k=5$, yang mengindikasikan pemisahan kluster memadai.

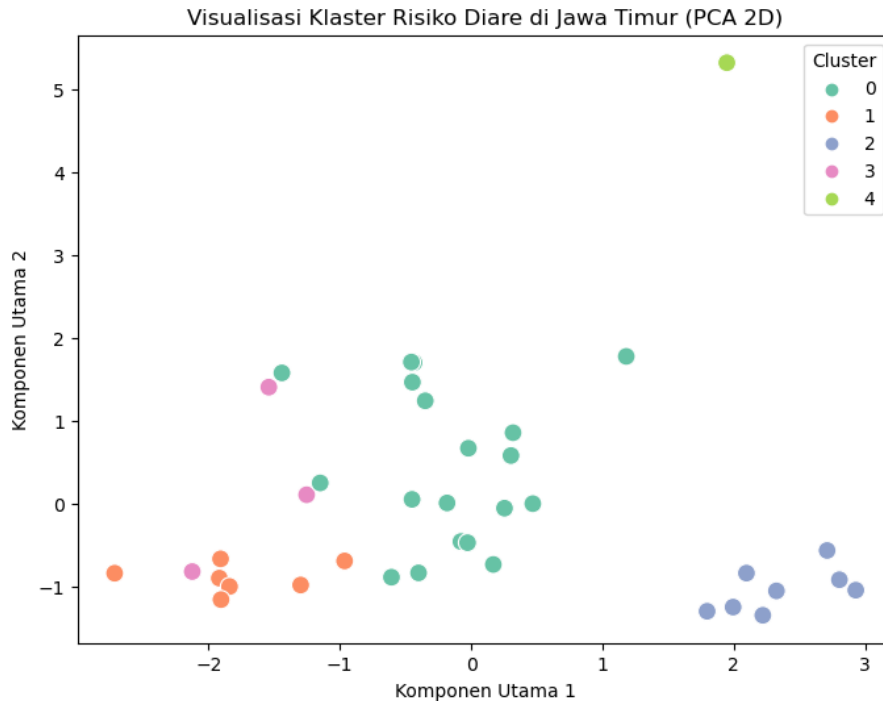
Tabel 3. Hasil Evaluasi Penentuan Jumlah Kluster Optimal Berdasarkan Metode *K-Means*

Jumlah Kluster (k)	Inertia	<i>Silhouette Score</i>	<i>Calinski-Harabasz</i>	<i>Davies-Bouldin</i>
2	159.440	0.347	15.480	0.865
3	120.231	0.308	15.686	1.240
4	89.548	0.339	17.523	0.926
5	61.452	0.400	22.35	0.745
6	47.319	0.352	24.437	0.802
7	42.394	0.326	22.620	0.848
8	37.830	0.336	21.544	0.879

Merujuk pada Tabel 3, struktur kluster paling optimal teridentifikasi pada lima kelompok ($k=5$). Penetapan ini didukung oleh nilai *Silhouette Score* tertinggi (sebesar 0,400) dan nilai *Davies-Bouldin Index* terendah (0,745), kedua indikator tersebut menegaskan pemisahan kluster yang optimal. Selain itu, peningkatan *Calinski-Harabasz Index* hingga 22,359 turut memvalidasi kualitas pemisahan (*separation*) antar kluster. Oleh karena itu, jumlah kluster optimal dalam analisis *K-Means* ditetapkan sebanyak lima kluster, karena konfigurasi ini menyediakan keseimbangan ideal antara kekompakan data di dalam kluster (*intra-cluster homogeneity*) dan perbedaan antar kluster (*inter-cluster heterogeneity*).

3.5. Visualisasi Hasil Kluster

Hasil klasterisasi ditampilkan melalui visualisasi, yang menampilkan hubungan antar klaster menggunakan pendekatan Analisis Komponen Utama Dua Dimensi (*Principal Component Analysis*), tersaji pada Gambar 4 sebagai berikut.



Gambar 4. Visualisasi Hasil Pengklasteran Risiko Kasus Diare di Jawa Timur

Gambar 4 menyajikan visualisasi klaster dengan Analisis Komponen Utama Dua Dimensi (PCA) menampilkan pemisahan yang jelas antar kelompok wilayah di Jawa Timur. Pola sebaran yang dihasilkan membuktikan bahwa lima klaster hasil analisis *K-Means* mampu mengidentifikasi variasi tingkat risiko diare antarwilayah. Klaster 0 merepresentasikan daerah dengan karakteristik menengah, Klaster 1 dan 3 terindikasi memiliki risiko tinggi, Klaster 2 tergolong rendah, sedangkan Klaster 4 menampilkan karakteristik ekstrem. Secara keseluruhan, pemisahan klaster yang jelas menunjukkan keberhasilan metode *K-Means* dalam menggambarkan heterogenitas risiko penyakit diare di Jawa Timur.

3.6. Hasil Klasterisasi Kabupaten/Kota Berdasarkan Risiko Penyakit Diare di Jawa Timur

Melalui Metode *K-Means*, wilayah di Provinsi Jawa Timur telah berhasil dikelompokkan menjadi lima kategori utama (Klaster 0–4), yang merefleksikan perbedaan profil risiko penyakit di setiap wilayah yang telah disajikan secara rinci di Tabel 4.

Tabel 4. Hasil Klaster Kabupaten/Kota di Jawa Timur

Klaster	Kabupaten/Kota
Klaster 0	Ponorogo, Tulungagung, Blitar, Kediri, Malang, Lumajang, Jember, Banyuwangi, Pasuruan, Sidoarjo, Mojokerto, Jombang, Nganjuk, Madiun, Magetan, Ngawi, Bojonegoro, Tuban, Gresik
Klaster 1	Bondowoso, Situbondo, Probolinggo, Bangkalan, Sampang, Pamekasan, Sumenep

Klaster 2	Kota Kediri, Kota Blitar, Kota Malang, Kota Probolinggo, Kota Pasuruan, Kota Mojokerto, Kota Madiun, Kota Batu
Klaster 3	Pacitan, Trenggalek, Lamongan
Klaster 4	Kota Surabaya

Berdasarkan hasil klasterisasi menggunakan metode *K-Means*, wilayah di Provinsi Jawa Timur terbagi menjadi lima kelompok utama (Klaster 0-4) dengan karakteristik sosial dan kesehatan yang berbeda.

- 1) Klaster 0 mencakup 19 kabupaten/kota, didominasi wilayah bagian tengah dan selatan Jawa Timur seperti Malang, Kediri, Blitar, dan Sidoarjo. Daerah-daerah ini memiliki infrastruktur kesehatan yang relatif baik dan akses air bersih tinggi, tetapi masih mencatat jumlah kasus diare yang cukup besar, sehingga dikategorikan sebagai wilayah dengan risiko sedang.
- 2) Klaster 1 terdiri dari tujuh kabupaten di wilayah Tapal Kuda dan Pulau Madura, seperti Sampang, Bangkalan, dan Pamekasan. Klaster ini dicirikan oleh akses sanitasi yang rendah dan tingkat kemiskinan yang tinggi, sehingga termasuk dalam kategori risiko tinggi terhadap penyakit diare.
- 3) Klaster 2 mencakup delapan kota di Jawa Timur seperti Kota Malang, Kota Batu, dan Kota Mojokerto. Klaster ini memiliki akses sanitasi dan air minum layak tertinggi, tingkat kemiskinan rendah, serta fasilitas kesehatan yang memadai, sehingga dikategorikan sebagai wilayah risiko rendah.
- 4) Klaster 3 terdiri dari tiga kabupaten yaitu Pacitan, Trenggalek, dan Lamongan, yang memiliki akses sanitasi sedang, kemiskinan menengah, dan jumlah puskesmas moderat, menandakan risiko sedang hingga tinggi.
- 5) Klaster 4 hanya terdiri dari satu kota, yaitu Surabaya, yang menjadi klaster unik dengan jumlah kasus diare tertinggi di provinsi ini (22.327 kasus). Meskipun memiliki fasilitas kesehatan terbaik, kepadatan penduduk yang ekstrem menjadikan kota ini berisiko tinggi akibat faktor urbanisasi.

3.7. Interpretasi Hasil Klasterisasi Berdasarkan Rata-Rata Variabel

Untuk memahami karakteristik unik tiap kelompok wilayah, interpretasi hasil klasterisasi dilakukan dengan meninjau profil rerata variabel dari masing-masing klaster (Tabel 5), menggambarkan perbedaan risiko penyakit diare di Jawa Timur.

Tabel 5. Rerata Hasil Klaster Berdasarkan Variabel Kabupaten/Kota di Jawa Timur

	Klaster 0	Klaster 1	Klaster 2	Klaster 3	Klaster 4
x_1	5751.42	3539.57	1259.25	5627.67	22327.00
x_2	96.78	96.60	99.38	81.44	98.15
x_3	87.19	65.35	93.59	84.58	97.81
x_4	9.88	16.59	5.82	12.43	4.72
x_5	947.58	702.57	5033.50	594.67	8667.00
x_6	31.53	24.71	7.38	26.00	63.00

Berdasarkan hasil analisis *K-Means*, diperoleh lima klaster wilayah dengan karakteristik sosial dan kesehatan yang berbeda-beda sebagaimana ditunjukkan pada Tabel 5. Pola yang terbentuk mencerminkan variasi tingkat risiko penyakit diare di Jawa Timur berdasarkan kombinasi indikator akses air bersih, sanitasi, kepadatan penduduk, kemiskinan, dan ketersediaan fasilitas kesehatan.

- 1) Klaster 0 mencakup wilayah dengan kondisi sosial-ekonomi yang relatif baik, ditandai oleh tingginya akses air minum layak (96,78%) dan sanitasi (87,19%),

serta tingkat kemiskinan yang rendah (9,88%). Meskipun demikian, rata-rata kasus diare masih tinggi (5.751 kasus). Hal ini menunjukkan bahwa keberadaan infrastruktur dasar belum sepenuhnya menekan kejadian diare, yang kemungkinan dipengaruhi oleh perilaku hidup bersih dan faktor lingkungan mikro. Daerah-daerah dalam klaster ini dapat dikategorikan sebagai wilayah risiko sedang.

- 2) Klaster 1 menunjukkan karakteristik risiko tinggi, dengan akses sanitasi yang rendah (65,35%) dan tingkat kemiskinan yang tinggi (16,59%). Jumlah kasus diare juga relatif besar, mencapai 3.539 kasus, sedangkan jumlah puskesmas per kabupaten masih terbatas (24 unit rata-rata). Wilayah dalam klaster ini umumnya memiliki faktor sosial dan infrastruktur dasar yang menjadi kendala utama dalam pencegahan penyakit berbasis lingkungan.
- 3) Klaster 2 menggambarkan wilayah risiko rendah. Klaster ini memiliki akses air minum dan sanitasi hampir sempurna (99,38% dan 93,59%) serta tingkat kemiskinan terendah (5,82%). Meskipun kepadatan penduduk sangat tinggi (5.033 jiwa/km²), jumlah kasus diare justru paling rendah (1.259 kasus). Kondisi ini menunjukkan efektivitas layanan kesehatan dan perilaku higienitas masyarakat perkotaan yang lebih baik dibandingkan wilayah lain.
- 4) Klaster 3 mencerminkan wilayah risiko sedang hingga tinggi yang memiliki akses air minum masih tergolong rendah (81,44%), dengan kemiskinan moderat (12,43%) dan jumlah puskesmas yang terbatas (26 unit). Meski jumlah kasus diare relatif tinggi (5.627 kasus), kondisi ini lebih disebabkan oleh keterbatasan geografis serta akses pelayanan kesehatan di daerah pinggiran.
- 5) Klaster 4 merupakan klaster unik dan ekstrem. Wilayah ini mencatat jumlah kasus diare tertinggi di Jawa Timur (22.327 kasus) dengan kepadatan penduduk mencapai 8.667 jiwa/km². Meskipun indikator sosial dan fasilitas kesehatan menunjukkan nilai optimal (akses sanitasi 97,81% dan 63 puskesmas), tingginya kepadatan dan mobilitas penduduk menjadikan Surabaya tetap berada pada kategori risiko tinggi akibat faktor urbanisasi dan lingkungan padat.

Secara keseluruhan, pola klaster menunjukkan bahwa penyakit diare tidak semata-mata dipengaruhi oleh satu faktor, melainkan kombinasi antara kondisi sosial-ekonomi, lingkungan fisik, dan fasilitas kesehatan. Daerah yang tingkat kemiskinannya tinggi dan sanitasinya buruk biasanya punya risiko lebih besar, sedangkan daerah dengan infrastruktur dan layanan kesehatan yang baik cenderung punya angka penyakit yang lebih rendah.

3.8. Rekomendasi dan Implikasi Kebijakan

Hasil klasterisasi menunjukkan bahwa risiko penyakit diare di Provinsi Jawa Timur tidak tersebar secara merata, melainkan membentuk pola spasial yang berkaitan erat dengan kondisi sosial, ekonomi, dan infrastruktur kesehatan daerah. Oleh karena itu, langkah penanganan perlu dirancang secara spesifik berdasarkan klaster agar upaya pengendalian diare menjadi lebih efektif dan efisien.

- 1) Klaster 1 (Risiko Tinggi):

Pemerintah daerah perlu memprioritaskan peningkatan akses sanitasi dan air minum layak, disertai edukasi Perilaku Hidup Bersih dan Sehat (PHBS) kepada masyarakat. Penguatan program STBM (Sanitasi Total Berbasis Masyarakat) dan pemberdayaan tenaga kesehatan desa menjadi kunci dalam menurunkan angka kasus di wilayah ini.

- 2) Klaster 0 dan 3 (Risiko Sedang):

Meskipun memiliki infrastruktur yang relatif memadai, wilayah ini tetap mencatat kasus diare yang tinggi. Fokus kebijakan perlu diarahkan pada pemantauan kualitas lingkungan, manajemen limbah rumah tangga, serta peningkatan cakupan imunisasi rotavirus dan penyuluhan gizi untuk anak balita.

- 3) Klaster 2 (Risiko Rendah Wilayah Perkotaan):

Daerah ini menunjukkan efektivitas sistem kesehatan dan perilaku masyarakat yang baik. Pemerintah perlu menjaga keberlanjutan program pencegahan serta memperkuat sistem surveilans penyakit berbasis masyarakat agar deteksi dini kasus dapat dilakukan dengan cepat.

4) Klaster 4 (Surabaya Klaster Unik Urban):

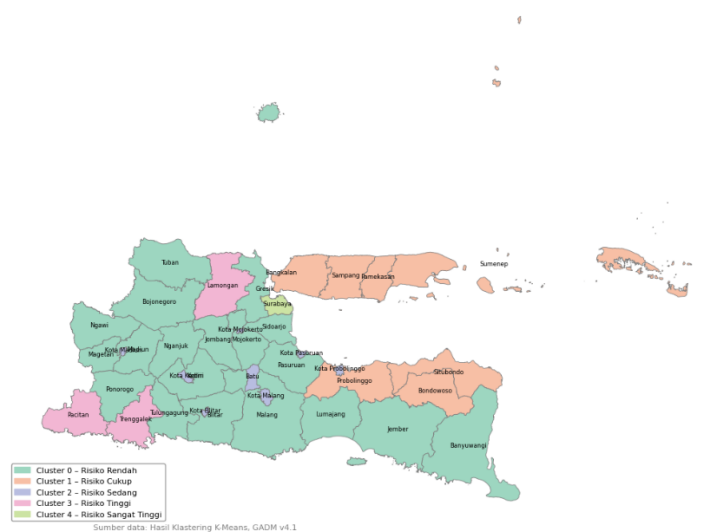
Kota ini menghadapi tantangan besar berupa kepadatan penduduk ekstrem dan mobilitas tinggi. Kebijakan yang perlu ditempuh mencakup pengelolaan permukiman padat, peningkatan infrastruktur sanitasi perkotaan, serta kolaborasi lintas sektor antara Dinas Kesehatan, Dinas PU, dan Dinas Lingkungan Hidup untuk mengendalikan faktor risiko lingkungan.

Secara umum, pendekatan berbasis klaster seperti yang digunakan dalam penelitian ini dapat mendukung pembuatan kebijakan kesehatan berbasis bukti (*evidence-based policy*). Melalui pemetaan spasial risiko, intervensi kesehatan masyarakat dapat diarahkan secara lebih tepat sasaran, sesuai dengan karakteristik sosial dan infrastruktur di masing-masing wilayah. Hasil ini diharapkan menjadi acuan bagi pemerintah daerah dalam menyusun strategi pencegahan diare yang adaptif, berkeadilan, dan berkelanjutan di seluruh kabupaten/kota di Jawa Timur.

3.9. Interpretasi Peta Sebaran Risiko Penyakit Diare di Provinsi Jawa Timur

Untuk memperjelas hasil analisis klasterisasi, dilakukan pemetaan spasial tingkat risiko penyakit diare di setiap kabupaten/kota di Provinsi Jawa Timur. Peta ini bertujuan untuk menggambarkan distribusi geografis antarwilayah berdasarkan hasil pengelompokan yang diperoleh dari metode *K-Means*.

Peta Klaster Risiko Penyakit Diare di Provinsi Jawa Timur



Gambar 5. Peta Sebaran Risiko Penyakit Diare di Provinsi Jawa Timur

Peta sebaran risiko penyakit diare di Provinsi Jawa Timur memperlihatkan pola spasial yang konsisten dengan hasil klasterisasi. Wilayah berisiko tinggi terkonsentrasi di kawasan Tapal Kuda dan Pulau Madura, sedangkan daerah dengan risiko rendah tersebar di bagian tengah dan selatan provinsi. Kota Surabaya muncul sebagai klaster tersendiri dengan tingkat risiko tertinggi akibat kepadatan penduduk yang ekstrem. Secara keseluruhan, peta ini menguatkan bahwa variasi risiko penyakit diare di Jawa Timur berkaitan erat dengan faktor sosial-ekonomi, kepadatan penduduk, dan kualitas infrastruktur kesehatan daerah.

IV. KESIMPULAN

Penelitian ini memetakan tingkat risiko penyakit diare di Provinsi Jawa Timur menggunakan algoritma *K-Means Clustering* dengan enam indikator sosial dan kesehatan. Hasil analisis menunjukkan terbentuknya lima klaster wilayah dengan tingkat risiko yang berbeda. Daerah Tapal Kuda dan Pulau Madura termasuk kategori risiko tinggi akibat rendahnya akses sanitasi dan tingginya kemiskinan, sedangkan wilayah perkotaan seperti Kota Malang dan Kota Mojokerto tergolong risiko rendah karena infrastruktur kesehatannya lebih baik. Kota Surabaya muncul sebagai klaster tersendiri dengan risiko tinggi akibat kepadatan penduduk ekstrem. Temuan ini menegaskan bahwa variasi risiko diare dipengaruhi oleh kondisi sosial-ekonomi, kepadatan, dan fasilitas kesehatan, sehingga kebijakan penanggulangan perlu disusun secara spesifik berbasis klaster agar lebih efektif dan berkeadilan.

UCAPAN TERIMA KASIH

Penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada Program Studi Sains Data, Universitas Pembangunan Nasional “Veteran” Jawa Timur atas dukungan fasilitas, arahan, dan kesempatan yang diberikan selama proses penelitian ini berlangsung. Ucapan terima kasih juga ditujukan kepada Badan Pusat Statistik (BPS) dan Dinas Kesehatan Provinsi Jawa Timur yang telah menyediakan data sekunder yang menjadi landasan utama dalam analisis penelitian ini. Penulis juga menghargai kontribusi rekan-rekan sejawat dan dosen pembimbing yang telah memberikan masukan berharga dalam penyusunan naskah ini. Selain itu, apresiasi disampaikan kepada panitia Seminar Nasional DISCOVER 2025 yang telah memberikan ruang bagi penulis untuk mempublikasikan hasil penelitian ini sebagai bagian dari upaya pengembangan ilmu pengetahuan di bidang sains data dan kesehatan masyarakat.

REFERENSI

- [1] Z. T. Ramadhani, L. Muniroh, and C. T. Abihail, “Hubungan Sumber Air Bersih, Kepadatan Hunian, dan Daerah Tempat Tinggal dengan Kejadian Diare pada Balita di Indonesia,” *Jurnal Ilmiah Universitas Batanghari Jambi*, vol. 23, no. 3, p. 3072, 2023, doi: 10.33087/jiubj.v23i3.4074.
- [2] Safira Alifia Husna and Nabila Vebiana Soviadi, “Distribusi Penyakit Diare Dan Determinan Dengan Pemetaan Wilayah Provinsi Jawa Barat Tahun 2020,” *Jurnal Kesehatan Komunitas Indonesia*, vol. 20, no. 2, pp. 136–146, 2024.
- [3] P. Vania and N. B. Sari, “Perbandingan Metode Elbow dan Silhouette untuk Penentuan Jumlah Klaster yang Optimal pada Clustering Produksi Padi menggunakan Algoritma K-Means,” *Jurnal Ilmiah Wahana Pendidikan*, vol. 9, no. November, pp. 684–694, 2023, doi: : <https://doi.org/10.5281/zenodo.10081332>.
- [4] A. P. Riani, A. Voutama, and T. Ridwan, “Penerapan K-Means Clustering Dalam Pengelompokan Hasil Belajar Peserta Didik Dengan Metode Elbow,” *J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD)*, vol. 6, no. 1, p. 164, 2023, doi: 10.53513/jsk.v6i1.7351.
- [5] H. Kurniawan, S. Defit, and Sumijan, “Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal Haris,”

JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST), vol. 1, no. 2, pp. 80–89, 2020.

- [6] P. P. Allorerung, A. Erna, M. Bagussahrir, and S. Alam, “Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit,” *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 178–191, 2024, doi: 10.14421/jiska.2024.9.3.178-191.
- [7] L. A. Sari, A. R. Hakim, and A. Rusgiyono, “Penggunaan Index Calinski-Harabasz Pada Clustering K-Medoids Algorithm Untuk Penggolongan Kabupaten/ Kota Di Provinsi Jawa Tengah Berdasarkan Karakteristik Penduduk Miskin,” *Jurnal Gaussian*, vol. 14, no. 1, pp. 179–187, 2025, doi: 10.14710/j.gauss.14.1.179-187.
- [8] L. E. Ekemeyong Awong and T. Zielinska, “Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification,” *Sensors*, vol. 23, no. 18, 2023, doi: 10.3390/s23187925.