Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

Segmentasi Kualitas Air Sungai Indonesia dengan Machine Learning untuk Penilaian Kelayakan Bahan Baku Minum

Shafira Febriani[1,*], Revi Firzatulloh[1], Rivaldo Nugraha[1], Anggi Cipta Lestari[2], Okyza Maherdi Prabowo[2], Yolanda Georgia Andriani[3]
[1]Program Studi Teknik Informatika, STMIK AMIK Bandung, Kota Bandung, Indonesia
[2] Program Studi Sistem Informasi, STMIK AMIK Bandung, Kota Bandung, Indonesia
[3] Program Studi Desain Komunikasi Visual, STMIK AMIK Bandung, Kota Bandung, Indonesia
Email: [1*]shafira@stmik-amikbandung.ac.id, [2]revif25@gmail.com, [3]rivaldo@stmik-amikbandung.ac.id, [4]anggi@stmik-amikbandung.ac.id, [5]okyza@stmik-amikbandung.ac.id, [6]yolanda@stmik-amikbandung.ac.id
Email Penulis Korespondensi: shafira@stmik-amikbandung.ac.id[*]
Submitted: **21/05/2026**; Accepted:**28/06/2026**; Published: **30/06/2026**

Abstrak–Sungai merupakan sumber daya air vital bagi kehidupan dan pembangunan di Indonesia, namun kualitasnya terus menurun akibat pencemaran domestik dan industri. Penelitian ini bertujuan untuk melakukan segmentasi kualitas air sungai di 38 provinsi di Indonesia berdasarkan Indeks Kualitas Air (IKA) serta parameter fisik, kimia, dan mikrobiologis tahun 2023. Segmentasi ini difungsikan untuk mendukung pemantauan dan penilaian kelayakan sungai sebagai bahan baku air minum. Metodologi penelitian mengikuti kerangka kerja Cross-Industry Standard Process for Data Mining (CRISP-DM) dengan menerapkan algoritma klasterisasi K-Means dan Hierarchical Clustering. Data yang dimanfaatkan bersumber dari Badan Pusat Statistik (BPS) yang mencakup parameter seperti pH, DO, BOD, COD, dan bakteri coliform. Meskipun evaluasi metrik internal (Silhouette Score, Davies-Bouldin Index, dan Dunn Index) menunjukkan bahwa konfigurasi dua klaster memberikan performa matematis terbaik , penelitian ini menetapkan konfigurasi empat klaster menggunakan Hierarchical Clustering. Langkah ini diambil secara sadar agar hasil pengelompokan tetap relevan secara praktis dan selaras dengan empat tingkat klasifikasi mutu air nasional (Kelas I–IV) serta skala penilaian IKA. Hasil segmentasi menghasilkan empat klaster utama yang menggambarkan tingkat risiko kualitas air: sangat tercemar, buruk hingga sedang, pencemaran ekstrem, dan risiko mikrobiologis tinggi. Temuan penelitian mengonfirmasi bahwa tidak ada provinsi yang memenuhi kriteria Kelas I (layak minum tanpa pengolahan), dengan mayoritas wilayah berada pada kategori sedang hingga buruk (Kelas III dan IV). Visualisasi melalui dashboard interaktif memperkuat interpretasi data untuk mendukung pengambilan kebijakan berbasis bukti. Penelitian ini membuktikan efektivitas metode klasterisasi dalam menyediakan informasi akurat bagi pengelolaan strategi kualitas air sungai secara nasional di Indonesia.

Kata Kunci: CRISP-DM; Kualitas Air; Hierarchical Clustering; Machine Learning; Segmentasi

Abstract– Rivers are vital water resources for life and development in Indonesia, yet their quality has declined due to domestic and industrial pollution. This study aims to segment river water quality across 38 provinces based on the Water Quality Index (WQI) and physical, chemical, and microbiological parameters from 2023. This segmentation serves to support assessment and monitoring of rivers as raw water sources for drinking water. The research methodology follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, applying K-Means and Hierarchical Clustering algorithms. The dataset, sourced from the Indonesian Central Bureau of Statistics (BPS), includes parameters such as pH, DO, BOD, COD, and coliform bacteria. Model evaluation using the Silhouette Score, Davies-Bouldin Index, and Dunn Index indicates that Hierarchical Clustering with a four-cluster configuration provides the most optimal and representative results. The segmentation identified four main clusters representing water quality risk levels: severely polluted, poor to moderate, extreme pollution, and high microbiological risk. The findings confirm that no province meets the Class I criteria (suitable for drinking without treatment), with the majority of regions falling into the moderate to poor categories (Class III and IV). Visualization through an interactive dashboard strengthens data interpretation to support evidence-based policymaking. This study demonstrates the effectiveness of clustering methods in providing accurate information for national river water quality management in Indonesia.

Keywords: CRISP-DM; Water Quality; Hierarchical Clustering; Machine Learning; Segmentation

1. PENDAHULUAN

Air merupakan sumber daya vital bagi keberlanjutan hidup manusia di berbagai sektor seperti konsumsi, sanitasi, pertanian, dan industri. Namun, kontaminasi mikrobiologis akibat tinja tetap menjadi ancaman terbesar bagi keamanan air minum dunia. Pada tahun 2022, sekitar 1,7 miliar penduduk global masih mengonsumsi air yang terkontaminasi tinja, yang berpotensi menyebarkan penyakit berbahaya seperti diare, kolera, disentri, tifus, hingga polio, dengan estimasi 505.000 kematian akibat diare setiap tahunnya [1]. Di Indonesia, pertumbuhan penduduk yang pesat dan urbanisasi memperparah tekanan terhadap ketersediaan air bersih [2]. Sungai yang memegang peran sentral sebagai penyedia air baku domestik kini menghadapi tantangan serius akibat pencemaran limbah domestik [3], yang tidak hanya menurunkan kelayakan air baku tetapi juga merusak keseimbangan ekosistem perairan.

Sebagai contoh, Sungai Krukut di DKI Jakarta mengalami penurunan kualitas yang signifikan akibat beban limbah domestik [4]. Data Kementerian Lingkungan Hidup dan Kehutanan (KLHK) dari 7.212 titik pemantauan sepanjang periode 2015–2023 menunjukkan bahwa mutu air sungai nasional pada tahun 2023 didominasi oleh kategori cemar ringan (66%), diikuti status baik atau kelas dua (21%), cemar sedang (13%), dan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

cemar berat (1%) [5]. Kendati pemerintah telah menerbitkan regulasi seperti Peraturan Pemerintah Nomor 22 Tahun 2021 tentang Penyelenggaraan Perlindungan dan Pengelolaan Lingkungan Hidup, implementasinya di lapangan masih belum optimal. Oleh karena itu, diperlukan pendekatan pemantauan baru yang adaptif dan berbasis teknologi untuk memperkuat pengawasan serta proses pengambilan keputusan.

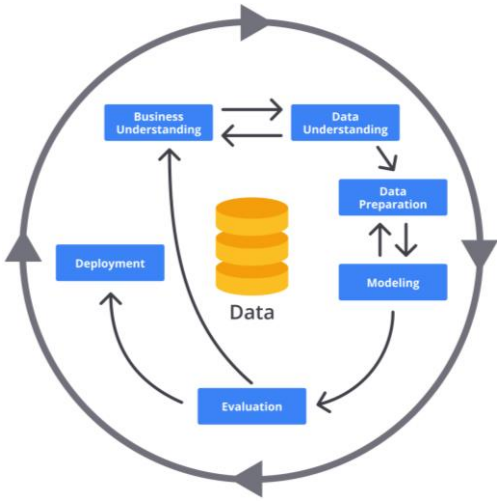
Machine learning hadir sebagai solusi analitik untuk melakukan segmentasi dan klasifikasi kualitas air secara akurat. Beberapa penelitian terdahulu telah membuktikan efektivitas metode ini, seperti pengelompokan sungai di Kota Semarang menjadi empat klaster menggunakan Kernel K-Means Clustering [6], serta identifikasi lima klaster ideal menggunakan Hierarchical Clustering berdasarkan parameter fisik dan kimia [7]. Selain itu, kajian Ahmed et al. [8] menunjukkan bahwa metode machine learning mampu meningkatkan akurasi prediksi hubungan nonlinier antarvariabel kualitas air. Pemodelan spasial-temporal kualitas air juga dapat membantu memetakan zona pencemaran potensial pada wilayah yang luas, sehingga sangat relevan untuk analisis kualitas air berskala nasional [9]. Namun, studi-studi tersebut umumnya masih bersifat lokal dan belum mengintegrasikan Indeks Kualitas Air (IKA) makro tingkat provinsi dengan parameter kualitas air sungai spesifik di ibu kota provinsi secara nasional.

Untuk mengisi celah riset tersebut, penelitian ini mengusulkan pendekatan segmentasi kualitas air sungai nasional tahun 2023 secara komprehensif. Algoritma K-Means diterapkan untuk menghasilkan partisi awal objek secara cepat, sedangkan Hierarchical Clustering digunakan untuk membangun struktur hubungan hierarkis dan visualisasi dendrogram yang mendalam. Berbeda dengan studi sebelumnya yang bertumpu pada satu atau dua metrik, validasi kualitas klasterisasi dalam penelitian ini dilakukan secara ketat menggunakan kombinasi tiga metrik internal sekaligus, yaitu Silhouette Score, Davies-Bouldin Index (DBI), dan Dunn Index.

Secara spesifik, penelitian ini dirumuskan untuk menjawab masalah sekaligus mencapai tiga tujuan utama, yaitu dengan melakukan segmentasi kualitas air sungai di 38 provinsi Indonesia secara menyeluruh dengan menggabungkan data IKA provinsi dan parameter kualitas air baku kota, membandingkan performa serta menentukan algoritma terbaik antara K-Means dan Hierarchical Clustering berdasarkan metrik evaluasi internal, dan menganalisis dan menyajikan hasil segmentasi ke dalam dashboard interaktif guna menghasilkan wawasan berbasis data untuk mendukung kebijakan pengelolaan air nasional yang lebih efektif.

2. METODOLOGI PENELITIAN

Dalam penelitian ini, digunakan pendekatan sistematis berbasis data mining untuk menganalisis dan melakukan segmentasi kualitas air sungai di Indonesia. Metodologi yang digunakan adalah Cross-lndustry Standard Process for Data Mining (CRISP-DM).



Gambar 1. Diagram Alur CRISP-DM Penelitian

2.1 Business Understanding

Tahap ini berfokus pada pemahaman masalah yang ingin dipecahkan melalui data mining, termasuk interaksi dengan pemangku kepentingan untuk mendefinisikan tujuan, target, dan kendala yang relevan [10]. Permasalahan utama yang diangkat adalah menurunnya kualitas air sungai di Indonesia, terutama akibat limbah domestik yang menyumbang lebih dari 75% pencemaran. Walaupun regulasi seperti PP No. 22 Tahun 2021 telah diterbitkan, implementasinya masih belum optimal sehingga pemantauan mutu air kurang efektif. Penelitian sebelumnya umumnya bersifat lokal, sehingga belum ada gambaran komprehensif di tingkat nasional dengan menggabungkan indeks kualitas air provinsi dan parameter sungai ibu kota provinsi. Penelitian ini bertujuan melakukan segmentasi kualitas air sungai nasional menggunakan data tahun 2023 dengan algoritma K-Means dan Hierarchical Clustering.Segmentasi ini diharapkan mampu mengelompokkan wilayah berdasarkan kemiripan karakteristik

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

kualitas air, sekaligus menyajikan visualisasi eksploratif untuk mendukung kebijakan berbasis data. Keberhasilan model diukur dengan Silhouette Score, Davies-Bouldin Index, dan Dunn Index, serta ditinjau dari kualitas visualisasi dan interpretasi hasil.

2.2 Data Understanding

Tahap ini berfokus pada pengumpulan dan eksplorasi dataset sekunder dari Badan Pusat Statistik (BPS) untuk memahami karakteristik dasar data [10]. Proses meliputi pemeriksaan struktur (jumlah observasi, variabel, dan tipe data), evaluasi kualitas data (nilai hilang dan potensi bias), analisis statistik deskriptif (rata-rata, median, standar deviasi, distribusi), serta Exploratory Data Analysis {EDA) melalui visualisasi distribusi, deteksi outlier, dan analisis korelasi antarvariabel. Tahap ini penting untuk mengidentifikasi potensi permasalahan data sena memberikan insight awal yang menjadi dasar persiapan dan pemodelan data. Dataset sekunder diperoleh secara resmi dari publikasi Badan Pusat Statistik (BPS) Indonesia mengenai Statistik Lingkungan Hidup Indonesia. Data aktual yang dianalisis mencakup kondisi kualitas air sungai di ibu kota 38 provinsi di Indonesia pada tahun periode 2023. Atribut input dalam eksperimen ini terdiri atas 12 parameter kualitas air, yaitu derajat keasaman (pH), suhu, Total Dissolved Solids (TDS), Total Suspended Solids (TSS), Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), nitrat (NO3), amonia (NH3), sulfat (SO4), Fecal Coliform, dan Total Coliform, serta dilengkapi dengan nilai makro Indeks Kualitas Air (IKA) 2023.

2.3 Data Preparation

Tahap ini memastikan data siap digunakan dengan kualitas memadai [10]. Proses mencakup imputasi nilai hilang menggunakan pendekatan berbeda untuk variabel kategorikal (“Tidak ada”) dan numerik (rata-rata atau median berbasis spasial dari provinsi tetangga), yang dilakukan iteratif untuk mengurangi kekosongan. Outlier tidak dihapus karena dianggap merepresentasikan kondisi lingkungan yang penting. Data kemudian dinormalisasi menggunakan Robust Scaler agar lebih tahan terhadap outlier [11]. Jumlah klaster ditetapkan empat, menyesuaikan klasifikasi mutu air nasional (Kelas I-IV) serta sistem penilaian Indeks Kualitas Air (IKA). Nilai kosong pada variabel numerik diimputasi menggunakan pendekatan berbasis spasial dari wilayah provinsi tetangga terdekat secara iteratif sebanyak tiga kali. Atribut dengan sebaran cenderung normal seperti pH, suhu, dan IKA diimputasi berbasis rata-rata (mean) , sedangkan parameter dengan sebaran miring (skewed) atau rentan pencilan ekstrem seperti TDS, BOD, COD, senyawa nitrogen/sulfat, dan coliform diimputasi berbasis nilai tengah (median). Nilai pencilan tidak dihapus melainkan dipertahankan demi menjaga representasi anomali kondisi lingkungan yang riil. Selanjutnya, seluruh fitur dinormalisasi menggunakan metode Robust Scaler yang mengandalkan nilai median dan Interquartile Range (IQR) agar parameter berbobot nilai besar tidak mendominasi model klasterisasi.

2.4 Modelling

Tahap pemodelan merupakan inti analisis dalam penelitian ini, dimana segmentasi kualitas air dilakukan menggunakan pendekatan unsupervised learning. Dua algoritma yang digunakan adlah K-Means Clustering dan Hierarchical Clustering, dengan tujuan mengelompokkan provinsi di Indonesia berdasarkan kemiripan parameter kualitas air. Segmentasi objek diuji secara paralel menggunakan dua algoritma unsupervised learning, yaitu K-Means Clustering dan Agglomerative Hierarchical Clustering. Berdasarkan regulasi baku mutu air nasional (PP No. 22 Tahun 2021) yang mengklasifikasikan air ke dalam empat tingkat (Kelas I hingga Kelas IV), jumlah klaster target (K) ditetapkan secara konstan sebanyak empat kelompok. Eksperimen pada pendekatan hierarkis menguji metode single, complete, dan average linkage , yang divalidasi lewat metrik Cophenetic Correlation Coefficient (CCC) guna mengukur akurasi dendrogram terhadap jarak asli data. Metode Average Linkage ditetapkan sebagai konfigurasi akhir karena terbukti menghasilkan nilai koefisien CCC tertinggi.

K-Means adalah algoritma klasterisasi non-hierarki yang membagi data ke dalam. K klaster berdasrkan kemiripan karakteristik antarobjek [12]. Langkah-langkah K-Means adalah sebagai berikut [13]:

1. Menentukan jumlah klaster (k) menggunakan metode Elbow.
2. Memilih centroid awal secara acak.
3. Menghitung jarak tiap data ke centroid menggunakan rumus jarak Euclidean:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{ik})^2} \qquad (1)$$

Dimana:
d_{ij} = jarak antara objek ke-i dan objek ke-j
p = jumlah klaster
x_{ik} = nilai data ke-i pada variabel ke-k
x_{jk} = nilai data ke-j pada variabel ke-k

4. Mengelompokkan data ke klaster dengan jarak terdekat.
5. Memperbarui posisi centroid dengan rata-rata data dalam klaster:

$$C_k = \left(\frac{1}{n_k}\right) \sum d_i \qquad (2)$$

Dimana:

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

d_{ij} = jarak antara objek ke-i dan objek ke-j
p = jumlah klaster
x_{ik} = nilai data ke-i pada variabel ke-k
x_{jk} = nilai data ke-j pada variabel ke-k

6. Mengulang langkah 3–5 hingga centroid stabil.

Hierarchical Clustering Adalah algoritma klasterisasi yang membentuk struktur hierarki dalam bentuk dendrogram, yang merepresentasikan tingkat kemiripan antar data [12]. Penelitian ini menggunakan pendekatan aglomeratif (*bottom-up*). Proses dimulai dari setiap data sebagai klaster individu yang kemudian digabungkan secara bertahap berdasarkan kedekatan jarak. Untuk menentukan metode penggabungan antar klaster (*linkage method*), dilakukan pengujian terhadap *single linkage*, *average linkage*, dan *complete linkage*. Evaluasi menggunakan *Cophenetic Correlation Coefficient (CCC)* menunjukkan bahwa *average linkage* memberikan hasil terbaik dengan nilai *CCC* = 0,9772, lebih tinggi dibandingkan *complete linkage* (0,9581) dan *single linkage* (0,9499). Oleh karena itu, *average linkage* dipilih sebagai metode utama. Rumus *average linkage* [14]:

$$d_{(ij)k} = \frac{\sum_a \sum_b d_{ab}}{N_{IJ} N_K} \quad (3)$$

Dimana:
d_{ab} = jarak antara objek i pada klaster IJ dan objek b pada klaster K
N_{IJ} = jumlah item pada klaster IJ
N_K = jumlah klaster IJ dan K

Untuk mengukur keakuratan struktur dendrogram digunakan *Cophenetic Correlation Coefficient* (*CCC*)[15]:

$$r_{coph} = \frac{\sum_{i<k}(d_{ik} - \bar{d})(d_{Cik} - \bar{d}_C)}{\sqrt{[\sum_{i<k}(d_{ik} - \bar{d})^2][\sum_{i<k}(d_{Cik} - \bar{d}_C)^2]}} \quad (4)$$

Dimana:
r_{coph} = *Cophenetic Correlation Coefficient*
d_{ik} = Jarak *Euclidean* objek ke- i dan ke- k
$\bar{d}$ = Rata-rata d_{ik}
d_{Cik} = Jarak *Cophenetic* objek ke- i dan ke- k
$\bar{d}_C$ = *Rata-rata* d_{Cik}

Langkah kerja algoritma *Hierarchical Clustering* meliputi setiap data diperlakukan sebagai klaster tunggal, menghitung jarak antar data, menggabungkan klaster terdekat, menghitung jarak baru antar klaster, dan membangun *dendrogram* hingga seluruh data tergabung, kemudian memotong *dendrogram* pada level tertentu untuk menentukan jumlah klaster optimal [14].

2.5 Evaluasi

Evaluasi model *clustering* bertujuan menilai kualitas hasil pengelompokan, terutama karena metode ini tidak memiliki label acuan. Beberapa metrik yang dimanfaatkan penggunaannya pada penelitian ini meliputi:

1. *Silhouette Coefficient*

Metrik ini mengukur sejauh mana sebuah data cocok berada dalam klasternya dibandingkan dengan klaster lain. Nilai berkisar dari -1 hingga 1, dengan nilai lebih tinggi menandakan pengelompokan lebih baik[16]. Rumusnya[13]:

$$S_i = \frac{b_i - a_i}{max(a_i - b_i)} \quad (5)$$

Dimana:
S_i = skor *silhouette* untuk data ke-i
a_i = rata-rata jarak antara i dan semua titik dalam klasternya sendiri

2. *Davies-Bouldin Index* (*DBI*)

Indeks ini menilai rasio antara dispersi dalam klaster (*intra-cluster*) dengan jarak antar klaster (*inter-cluster*). Semakin kecil nilai *DBI*, semakin baik kualitas klaster. Rumusnya [16] :

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

Dimana:
k = jumlah *cluster*
$R_{i,j}$ = rasio perbandungan antara *cluster* ke-i dan *cluster* ke-j

3. *Dunn Index*

Indeks ini digunakan untuk mengukur sejauh mana klaster terpisah dengan jelas satu sama lain. Nilai yang tinggi menunjukkan klaster memiliki pemisahan yang baik sekaligus kohesi internal yang kuat[17]. Rumusnya [13]:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \quad max_{i \neq j}(R_{i,j}) \quad (7)$$

Dimana:

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

$d(C_i, C_j)$ = jarak antara pusat klaster C_i dan C_j
$d(C_k)$ = jarak maksimum antara titik dalam klaster C_k

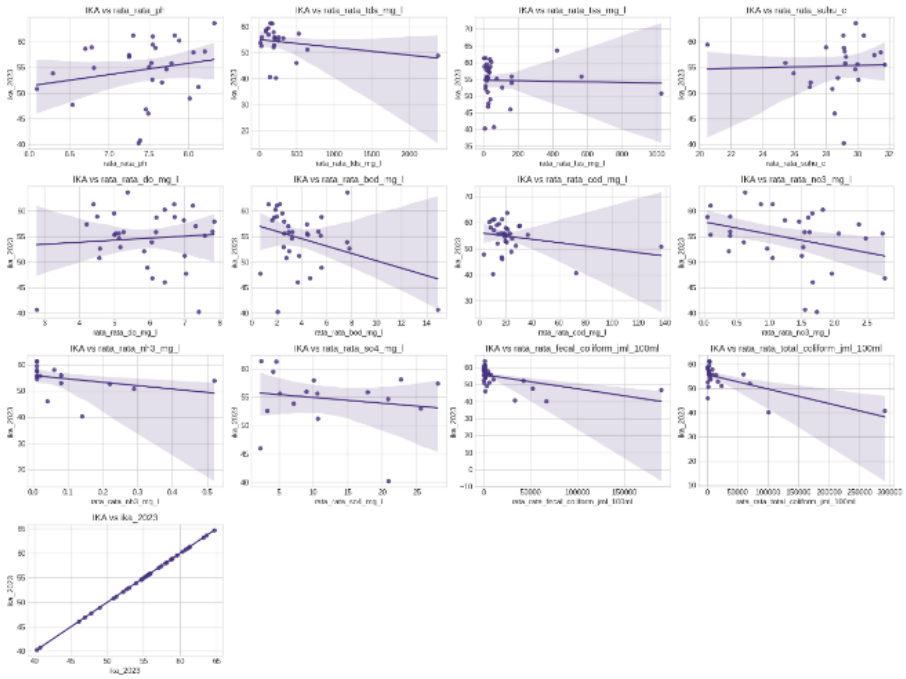
3. HASIL DAN PEMBAHASAN

3.1 Karakteristik Data dan Hasil Preprocessing

Bagian ini menyajikan hasil analisis awal terhadap data yang dimanfaatkan penggunaannya pada penelitian. Proses ini mencakup pemahaman terhadap struktur dan isi dataset serta tahapan pra-pemrosesan yang dilakukan untuk memastikan bahwa data dalam kondisi layak digunakan dalam proses pemodelan. Seluruh langkah yang dijalankan bertujuan untuk meningkatkan kualitas data sehingga hasil yang diperoleh dari tahap analisis berikutnya menjadi valid dan dapat diandalkan.

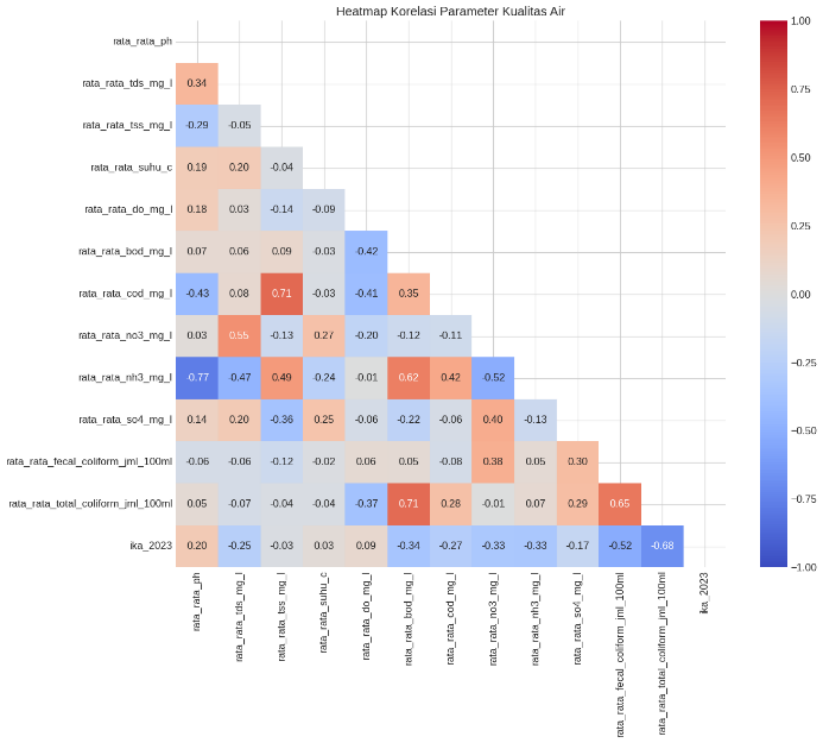
1. Struktur Dataset
   Langkah pertama adalah mengevaluasi struktur dataset yang dimanfaatkan penggunaannya pada penelitian. Dataset terdiri dari 38 entri yang merepresentasikan kota-kota di berbagai provinsi, dengan total 16 variabel yang mencakup parameter fisik, kimia, dan mikrobiologis kualitas air sungai. Terdapat tiga kolom kategorikal (provinsi, kota, dan sungai) serta 13 kolom numerik, meliputi pH, suhu, TDS, TSS, DO, BOD, COD, serta konsentrasi senyawa seperti NO_3, NH_3, dan SO_4. Selain itu ada Indeks Kualitas Air (IKA) tahun 2023 untuk variabel tambahan, tersedia lengkap untuk seluruh entri, sehingga dapat dimanfaatkan secara optimal dalam pemodelan. Meskipun demikian, hasil eksplorasi menunjukkan adanya nilai hilang pada beberapa parameter penting. Kolom sungai memiliki lima nilai kosong, sementara variabel numerik seperti NH_3 dan SO_4 masing-masing kehilangan 21 nilai, serta TDS, suhu, dan total coliform yang masing-masing kehilangan 15 nilai. Kehadiran data hilang ini menegaskan perlunya pra-pemrosesan sebelum dilakukan analisis lebih lanjut.
2. Deskripsi Statistik Dataset
   Analisis statistik deskriptif dijalankan guna memperoleh gambaran umum distribusi dan sebaran data. Hasilnya menunjukkan bahwa pH air sungai berkisar antara 6,9 hingga 8,32 dengan rata-rata 7,43, mencerminkan kondisi air netral hingga sedikit basa. Parameter TDS dan TSS menunjukkan variasi yang sangat tinggi, dengan nilai maksimum masing-masing mencapai 2.380,67 mg/L dan 1.028,78 mg/L, mengindikasikan potensi pencemaran bahan terlarut dan sedimen. Kandungan oksigen terlarut (DO) rata-rata sebesar 5,90 mg/L tergolong cukup baik, meskipun nilai minimum 2,78 mg/L menandakan adanya potensi kondisi anaerobik di lokasi tertentu. Tingkat pencemaran organik tercermin dari nilai BOD (3,92 mg/L) dan COD (23,47 mg/L) dengan variasi yang tinggi, menandakan adanya beban pencemar organik maupun kimia. Konsentrasi nitrat (NO_3) masih dalam batas aman, sedangkan amonia (NH_3) menunjukkan nilai maksimum 0,52 mg/L yang perlu mendapat perhatian. Pencemaran mikrobiologis tampak signifikan, dengan rata-rata fecal coliform mencapai 13.335,81 jml/100ml dan total coliform 26.980,98 jml/100ml, menunjukkan adanya kontaminasi biologis serius di beberapa sungai. Nilai rata-rata IKA tahun 2023 adalah 55,29, yang mengindikasikan bahwa kualitas air sungai di Indonesia secara umum berada pada kategori sedang hingga cukup baik, meskipun belum memenuhi standar ideal.
3. Hasil *Explatory Data Analysis* (*EDA*)
   Exploratory Data Analysis dilakukan untuk memperdalam pemahaman terhadap karakteristik data. Visualisasi kategori IKA per provinsi menunjukkan bahwa mayoritas provinsi (32 dari 38) berada dalam kategori sedang (Kelas III), dengan Papua Tengah mencatat nilai tertinggi (64,67). Sebaliknya, enam provinsi, termasuk DI Yogyakarta (40,28) dan DKI Jakarta (40,76), masuk kategori buruk (Kelas IV). Temuan ini menegaskan adanya disparitas kualitas air antarwilayah yang perlu ditangani secara khusus. Analisis hubungan antara IKA dan parameter utama menunjukkan pola korelasi yang signifikan. Parameter pencemar organik dan biologis, seperti BOD, COD, NH_3, serta coliform, memiliki korelasi negatif terhadap IKA, yang berarti peningkatan konsentrasi senyawa tersebut berasosiasi dengan penurunan kualitas air. Sebaliknya, DO dan pH menunjukkan hubungan positif terhadap IKA, yang menandakan bahwa ketersediaan oksigen terlarut dan kestabilan pH berkontribusi pada kualitas air yang lebih baik, sebagaimana diperlihatkan pada Gambar. 1.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824



Gambar 2. Hubungan IKA dengan Parameter Kunci

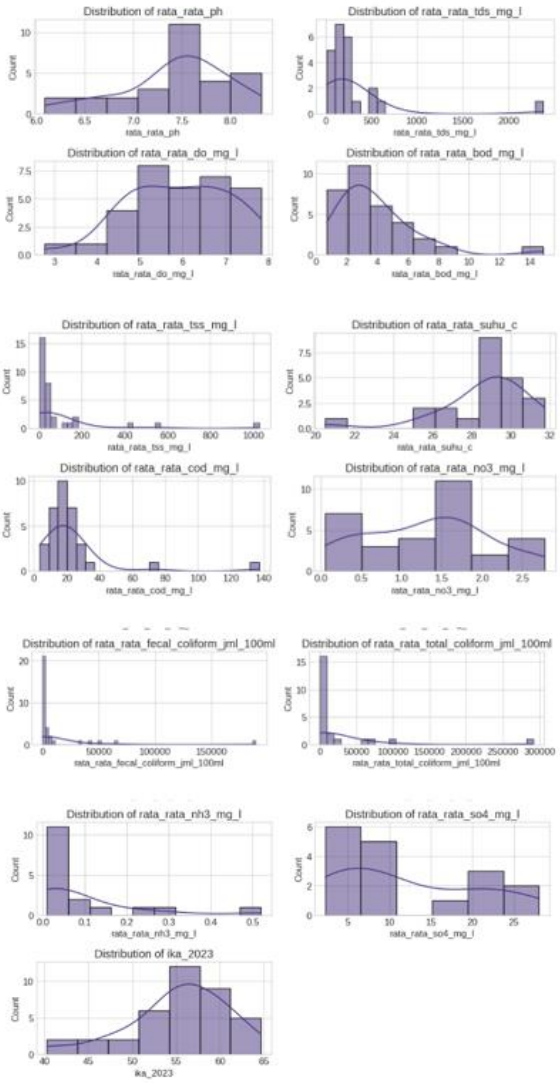
Correlation matrix juga menegaskan bahwa total coliform memiliki pengaruh paling dominan terhadap penurunan kualitas air (r = -0,68), diikuti fecal coliform dan amonia. Korelasi antarparameter seperti COD dan BOD (r = 0,71) menunjukkan keterkaitan erat antara pencemar organik dan kimia, sebagaimana diperlihatkan pada Gambar. 2.



Gambar 3. Korelasi Kolom Numerik

Distribusi parameter numerik juga memperlihatkan keberadaan sejumlah *outlier*, terutama pada variabel TDS, TSS, coliform, serta BOD dan COD. *Outlier* tersebut dianggap penting untuk dipertahankan karena merepresentasikan kondisi ekstrem kualitas air yang relevan secara lingkungan, sebagaimana diperlihatkan pada Gambar. 3.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824



Gambar 4. Distribusi Kolom Numerik

4. Hasil Persiapan Data

Proses pra-pemrosesan data dilakukan secara sistematis untuk memastikan kualitas dataset. Pertama, seluruh nilai hilang berhasil diimputasi melalui pendekatan berbasis spasial dengan mempertimbangkan kesamaan antarprovinsi, sehingga jumlah nilai hilang berkurang hingga nol setelah tiga iterasi. Kedua, analisis *outlier* menunjukkan bahwa sebagian besar variabel mengandung nilai ekstrem, namun nilai-nilai tersebut tetap dipertahankan untuk menjaga representasi kondisi nyata lapangan, sebagaimana diperlihatkan pada Tabel 1.

Tabel 1. jumlah outlier tiap variabel setelah di imputasi nilai yang hilang

No	Nama Variabel	Jumlah *Outlier*
1	rata_rata_ph	3
2	rata_rata_tds_mg_l	5
3	rata_rata_tss_mg_l	7
4	rata_rata_suhu_c	2
5	rata_rata_do_mg_l	0
6	rata_rata_bod_mg_l	3
7	rata_rata_cod_mg_l	4
8	rata_rata_no3_mg_l	0
9	rata_rata_nh3_mg_l	1
10	rata_rata_so4_mg_l	0
11	rata_rata_fecal_coliform_jml_100ml	5
12	rata_rata_total_coliform_jml_100ml	8
13	ika_2023	2

Ketiga, proses normalisasi menggunakan metode *Robust Scaler* berhasil menyetarakan skala antarvariabel tanpa menghilangkan pengaruh nilai ekstrem. Hasil transformasi menunjukkan bahwa seluruh fitur

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

memiliki skala yang relatif seimbang, sehingga dapat berkontribusi proporsional dalam proses klasterisasi, sebagaimana diperlihatkan pada Gambar. 4.

provinsi	kota	sungai	rata_rata_ph	rata_rata_tds_mg_l	rata_rata_tss_mg_l	rata_rata_suhu_c	rata_rata_do_mg_l	rata_rata_bod_mg_l	rata_rata_cod_mg_l
Aceh	Banda Aceh	Krueng Sabee	-0.564972	-0.357213	-0.619583	0.095628	0.610947	-0.715640	-0.767748
Sumatera Utara	Medan	Deli	0.745763	-0.357213	0.446505	-0.095628	0.445266	-0.454976	-1.218814
Sumatera Barat	Padang	Batang Ombilin	-0.677966	-0.580579	0.169001	0.183060	0.971893	0.028436	-0.802805
Riau	Pekanbaru	Batang Gangsal	-3.299435	-0.580579	26.314329	-0.286885	-0.649408	-0.028436	13.905930
Jambi	Jambi	Batanghari	-0.135593	2.657541	3.361215	-0.215847	0.445266	0.398104	-0.088811

provinsi	kota	sungai	rata_rata_no3_mg_l	rata_rata_nh3_mg_l	rata_rata_so4_mg_l	rata_rata_fecal_coliform_jml_100ml	rata_rata_total_coliform_jml_100ml	ika_2023
Aceh	Banda Aceh	Krueng Sabee	-0.593968	-0.076923	-0.569816	-0.172427	-0.065377	0.898058
Sumatera Utara	Medan	Deli	0.287703	1.000000	-0.569816	-0.285325	-0.127199	0.731392
Sumatera Barat	Padang	Batang Ombilin	-0.139211	1.115385	-0.581859	-0.210048	-0.139439	0.210356
Riau	Pekanbaru	Batang Gangsal	-0.426914	2.076923	-0.581859	-0.224021	-0.189021	-0.794498
Jambi	Jambi	Batanghari	0.148492	0.153846	-0.581859	-0.059530	-0.227393	-1.567961

Gambar 5. Hasil Normalisasi

3.2 Hasil Pemodelan dan Evaluasi

Pada tahap ini disajikan hasil penerapan dan evaluasi dua metode klasterisasi, yaitu *K-Means* dan *Hierarchical Clustering* dengan pendekatan *average linkage*, terhadap data kualitas air sungai. Evaluasi dilakukan untuk menentukan kombinasi terbaik antara teknik pra-pemrosesan dan jumlah klaster, berdasarkan tiga metrik evaluasi internal yakni *Silhouette Score*, *Davies-Bouldin Index* (*DBI*), dan *Dunn Index*. Rekap hasil evaluasi untuk jumlah klaster 2 hingga 9 dari masing-masing model ditampilkan pada Tabel 2 dan Tabel 3.

1. Hasil Segmentasi Model *K-Means*
   Algoritma *K-Means* digunakan dengan teknik *Robust Scaler* untuk normalisasi data, tanpa penanganan khusus terhadap *outlier*. Evaluasi dilakukan pada jumlah klaster 2 hingga 9, berguna mengidentifikasi konfigurasi yang paling optimal. Performa model dihitung menggunakan tiga metrik internal. *Silhouette Score* menilai kedekatan suatu observasi dengan klasternya dibandingkan dengan klaster lain. *Davies-Bouldin Index* (*DBI*) mengukur rasio antara dispersi dalam klaster dengan jarak antar klaster, di mana nilai lebih rendah menunjukkan hasil yang lebih baik. Sementara itu, *Dunn Index* membandingkan jarak minimum antar klaster dengan diameter maksimum dalam klaster, sehingga nilai yang lebih tinggi memperlihatkan pemisahan antar klaster yang lebih baik. Hasil evaluasi diperlihatkan melalui Tabel 2.

Tabel 2. Hasil Evaluasi *K-Means*

Jumlah Klaster	*Silhouette Score*	*Davies-Bouldin Index*	*Dunn Index*
2	0.5683	1.0748	0.1601
3	0.5469	0.9475	0.1601
4	0.5528	0.5703	0.3380
5	0.5238	0.5669	0.2804
6	0.5088	0.4801	0.3891
7	0.5051	0.4941	0.5207
8	0.4838	0.4129	0.5993
9	0.2945	0.5387	0.2991

Berdasarkan hasil tersebut, konfigurasi terbaik dicapai pada jumlah delapan klaster, dengan *Silhouette Score* 0.4838, *DBI* 0.4129, dan *Dunn Index* 0.5993. Konfigurasi ini memberikan keseimbangan terbaik antara kepadatan internal dan pemisahan antar klaster.

2. Hasil Segmentasi Model *Hierarchical Clustering*
   Metode *Hierarchical Clustering* diterapkan menggunakan *average linkage* dengan normalisasi data melalui *Robust Scaler*, tanpa penanganan *outlier*. Evaluasi dilakukan pada jumlah klaster 2 hingga 9. Hasil evaluasi diperlihatkan melalui Tabel 3.

Tabel 3. Hasil Evaluasi *Hierarchical Clustering*

Jumlah Klaster	*Silhouette Score*	*Davies-Bouldin Index*	*Dunn Index*
2	0.6822	0.2041	0.6377
3	0.6173	0.2112	0.5063
4	0.5702	0.3744	0.6168
5	0.5356	0.3423	0.4760
6	0.4921	0.4168	0.4551
7	0.4681	0.3346	0.4551
8	0.4611	0.3728	0.5809
9	0.4290	0.3606	0.7005

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

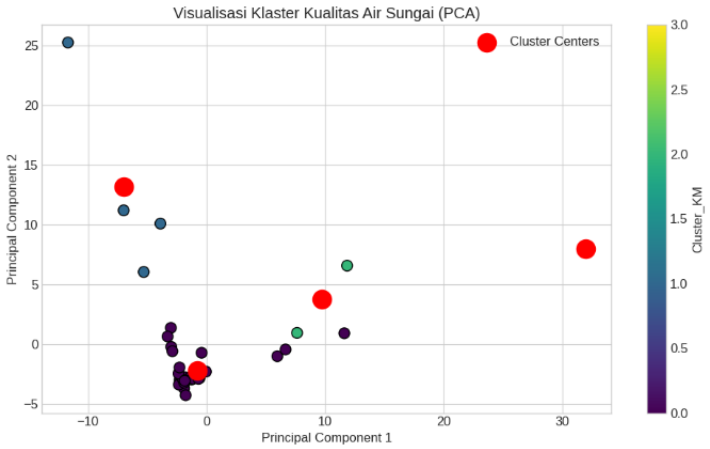
Hasil evaluasi memperlihatkan bahwasanya konfigurasi dengan jumlah dua klaster memberikan performa terbaik berdasarkan metrik evaluasi internal *Silhouette Score* sebesar 0.6822, *DBI* sebesar 0.2041, *Dunn Index* sebesar 0.6377. Dari hasil evaluasi kedua model diatas dapat disimpulkan bahwa hasil dari *Hierarchical Clustering* lebih unggul daripada *K-Means*. Namun, konfigurasi dengan dua klaster tersebut tidak digunakan karena tidak sesuai dengan tujuan penelitian yang mengacu pada klasifikasi mutu air nasional, yang terbagi menjadi empat kelas (Kelas I–IV). Selain itu, pendekatan ini juga diselaraskan dengan sistem penilaian Indeks Kualitas Air (IKA) yang menggunakan skala klasifikasi yang serupa. Oleh karena itu, dipilih konfigurasi dengan empat klaster untuk analisis lanjutan.

3. Perbandingan dan Pemilihan Model Terbaik
   Untuk konsistensi dengan sistem klasifikasi mutu air nasional, kedua metode dibandingkan dengan konfigurasi empat klaster. Hasil evaluasi ditampilkan pada Tabel 4.

Tabel 4. Evaluasi Model Pada Konfigurasi Empat Klaster

Model	*Silhouette Score*	*Davies-Bouldin Index*	*Dunn Index*
K-Means	0.5528	0.5703	0.3380
Hierarchical Clustering	0.5702	0.3744	0.6168

Hasil menunjukkan bahwa *Hierarchical Clustering* lebih unggul dibandingkan *K-Means*, dengan nilai *Silhouette Score* yang lebih tinggi, *DBI* yang lebih rendah, serta *Dunn Index* yang lebih tinggi. Hal ini juga didukung oleh hasil Tabel 3 yang menghasilkan *Hierarchical Clustering* dengan dua klaster juga lebih unggul dari *K-Means*.



Gambar 6. Visualisasi *Principal Component Analysis (PCA) K-Means*

Visualisasi hasil K-Means melalui reduksi dimensi *Principal Component Analysis (PCA)* pada Gambar. 5 menunjukkan adanya tumpang tindih antar klaster, sedangkan *Hierarchical Clustering* menghasilkan dendrogram yang memperlihatkan struktur pengelompokan lebih jelas. Dari segi *Cohesion* (Kekompakan Intra-Klaster), *Hierarchical Clustering* lebih kompak, dengan nilai *Dunn Index* 0.6168 yang menunjukkan jarak antar klaster cukup besar dibandingkan penyebaran dalam klaster. Dari segi *Separation* (Pemisahan Antar Klaster), nilai *DBI* 0.3744 yang lebih rendah juga menegaskan pemisahan antar klaster yang lebih baik dibanding *K-Means*. Keunggulan lain dari *Hierarchical Clustering* adalah kemampuannya mengungkap hubungan hierarkis antar provinsi, yang bermanfaat dalam kebijakan pengelolaan wilayah. Berdasarkan metrik evaluasi, kejelasan segmentasi, serta kesesuaiannya dengan sistem klasifikasi mutu air nasional, model terbaik yang dipilih adalah *Hierarchical Clustering* dengan empat klaster.

3.3 Pembahasan dan Analisis Hasil Segmentasi

Hasil analisis memperlihatkan bahwasanya metode *Hierarchical Clustering* (average linkage) memberikan performa lebih baik dibandingkan K-Means berdasarkan *Silhouette Score*, *Davies-Bouldin Index*, dan *Dunn Index*. Oleh karena itu, pembahasan lebih lanjut difokuskan pada segmentasi yang dihasilkan oleh metode tersebut. Dataset segmentasi kualitas air sungai di 38 provinsi Indonesia diperoleh dari gabungan data parameter fisik, kimia, dan mikrobiologis kualitas air sungai di ibu kota provinsi dengan Indeks Kualitas Air (IKA) tahun 2023. Dataset ini juga diperkaya dengan informasi geografis (*latitude* dan *longitude*) untuk mendukung analisis spasial. Tujuan segmentasi adalah menilai kesesuaian mutu air sungai terhadap baku mutu nasional serta potensi pemanfaatannya sebagai bahan baku air minum. Berdasarkan Peraturan Pemerintah No. 22 Tahun 2021, air sungai diklasifikasikan ke dalam empat kelas mutu:

1. Kelas I, layak sebagai bahan baku air minum tanpa pengolahan lanjut,
2. Kelas II, untuk rekreasi air, budidaya ikan air tawar, peternakan, dan irigasi pertanian,
3. Kelas III, untuk budidaya ikan, peternakan, dan pertanian,

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

4. Kelas IV, untuk pengairan tanaman atau pemanfaatan lain dengan kualitas rendah.

Penelitian ini menggunakan 12 parameter utama (temperatur, TDS, TSS, pH, BOD, COD, DO, sulfat, nitrat, amoniak, fecal coliform, total coliform) yang selaras dengan baku mutu nasional. Penelitian ini menggunakan 12 parameter utama (temperatur, TDS, TSS, pH, BOD, COD, DO, sulfat, nitrat, amoniak, fecal coliform, total coliform) yang selaras dengan baku mutu nasional. Rincian ambang batas tiap kelas mutu ditampilkan pada Tabel 5. Analisis kualitas air dilakukan dengan mengacu pada nilai ambang ini serta IKA, sehingga evaluasi kondisi setiap provinsi bersifat lebih komprehensif.

TABEL I
BAKU MUTU AIR SUNGAI BERDASARKAN KELAS MUTU NASIONAL

No	Parameter	Unit	Kelas 1	Kelas 2	Kelas 3	Kelas 4	Keterangan
1	Temperatur	°C	Dev 3	Dev 3	Dev 3	Dev 3	Perbedaan dengan suhu udara di atas permukaan air
2	Padatan terlarut total (TDS)	mg/L	1.000	1.000	1.000	2.000	Tidak berlaku untuk muara
3	Padatan tersuspensi total (TSS)	mg/L	40	50	100	400	
4	Derajat keasaman (pH)		6 - 9	6 - 9	6 - 9	6 - 9	Tidak berlaku untuk air gambut (berdasarkan kondisi alaminya)
5	Kebutuhan oksigen biokimiawi (BOD)	mg/L	2	3	6	12	
6	Kebutuhan oksigen kimiawi (COD)	mg/L	10	25	40	80	
7	Oksigen terlarut (DO)	mg/L	6	4	3	1	Batas Minimal
8	Sulfat (SO_4^2)	mg/L	300	300	300	400	
9	Nitrat (sebagai N)	mg/L	10	10	20	20	
10	Amoniak (sebagai N)	mg/L	0,1	0,2	0,5	-	
11	Fecal Coliform	MPN/ 100 mL	100	1.000	2.000	2.000	
12	Total Coliform	MPN/ 100 mL	1.000	5.000	10.000	10.000	

Hasil *Hierarchical Clustering* membagi 38 provinsi menjadi empat klaster yang mencerminkan variasi kondisi mutu air:

1. *Cluster* 1 Provinsi Sangat Tercermar
   Klaster pertama terdiri dari DKI Jakarta dan Banten. Wilayah ini menunjukkan kualitas air yang sangat buruk dengan nilai IKA masing-masing 40,76 dan 58,93, mengindikasikan masuk dalam kategori Kelas IV. Parameter pencemar utama di klaster ini adalah kadar DO yang rendah, BOD dan COD yang tinggi, serta angka coliform yang luar biasa besar DKI Jakarta memiliki total coliform sebesar 291.782 MPN/100 ml. Karakteristik ini menandakan air tidak layak untuk keperluan apapun tanpa pengolahan lanjutan dan berisiko besar terhadap kesehatan masyarakat.
2. *Cluster* 2 Provinsi dengan Kualitas Buruk hingga Sedang
   Klaster kedua mencakup 34 provinsi seperti Aceh, Sumatera Barat, Kalimantan Timur, Nusa Tenggara Barat, hingga Papua. Provinsi-provinsi ini memiliki nilai IKA bervariasi antara 46 hingga 64, dengan rata-rata di kisaran 54–60, yang tergolong sedang hingga buruk dan umumnya masuk Kelas III–IV. Meskipun pH dan suhu berada pada rentang normal, sebagian besar daerah memiliki kadar DO di bawah 6 mg/L, menunjukkan kondisi oksigen terlarut yang kurang mendukung ekosistem perairan. Selain itu, kadar BOD, COD, dan coliform cenderung tinggi. Meskipun tingkat pencemaran tidak separah klaster pertama, provinsi dalam klaster ini memerlukan upaya peningkatan pengelolaan limbah dan pengawasan mutu air secara berkala.
3. *Cluster* 3 Provinsi dengan Pencemaran Ekstrem
   Klaster ketiga hanya diwakili oleh Provinsi Riau yang memiliki nilai IKA sebesar 50,84. Provinsi ini tergolong sangat tercemar dengan karakteristik menonjol pada parameter TSS yang ekstrem, mencapai 1026,76 mg/L, serta nilai COD sebesar 137,04 mg/L yang jauh melebihi baku mutu kelas IV. Hal ini menunjukkan pencemaran partikel tersuspensi dan zat organik sangat tinggi, menjadikan air di wilayah ini tidak layak digunakan tanpa pengolahan serius.
4. *Cluster* 4 Provinsi dengan Risiko Mikrobiologis Tinggi
   Klaster keempat terdiri dari Provinsi Jawa Barat yang menunjukkan kualitas air dengan nilai IKA 46,87. Meskipun parameter fisik dan kimia seperti pH dan suhu tergolong baik, pencemaran mikrobiologis sangat tinggi, ditunjukkan oleh total coliform mencapai 157.391 MPN/100 ml. Provinsi ini masuk ke dalam klasifikasi Kelas IV, dan risiko terbesar berasal dari kontaminasi biologis, yang dapat menimbulkan dampak serius terhadap kesehatan masyarakat jika air digunakan tanpa disterilkan

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

Evaluasi parameter kualitas air di 38 provinsi mengungkapkan beberapa hal penting. Nilai pH di sebagian besar provinsi masih berada pada rentang normal (6–9), dan suhu air berkisar antara 26–31°C, yang masih sesuai dengan standar baku mutu kelas I–IV. Namun, kadar oksigen terlarut (DO) cenderung rendah di banyak wilayah (< 5 mg/L), mengindikasikan penurunan kualitas ekosistem akuatik. Parameter BOD dan COD di sebagian besar wilayah melebihi ambang batas baku mutu kelas II dan III, menunjukkan tingginya beban pencemar organik. TSS dan TDS yang tinggi ditemukan pada beberapa daerah seperti Riau dan Bengkulu, menunjukkan beban sedimen dan zat terlarut yang berat. Selain itu, pencemaran mikrobiologis menjadi isu utama yang merata di banyak provinsi, dengan nilai total coliform dan fecal coliform jauh melampaui baku mutu kelas IV.

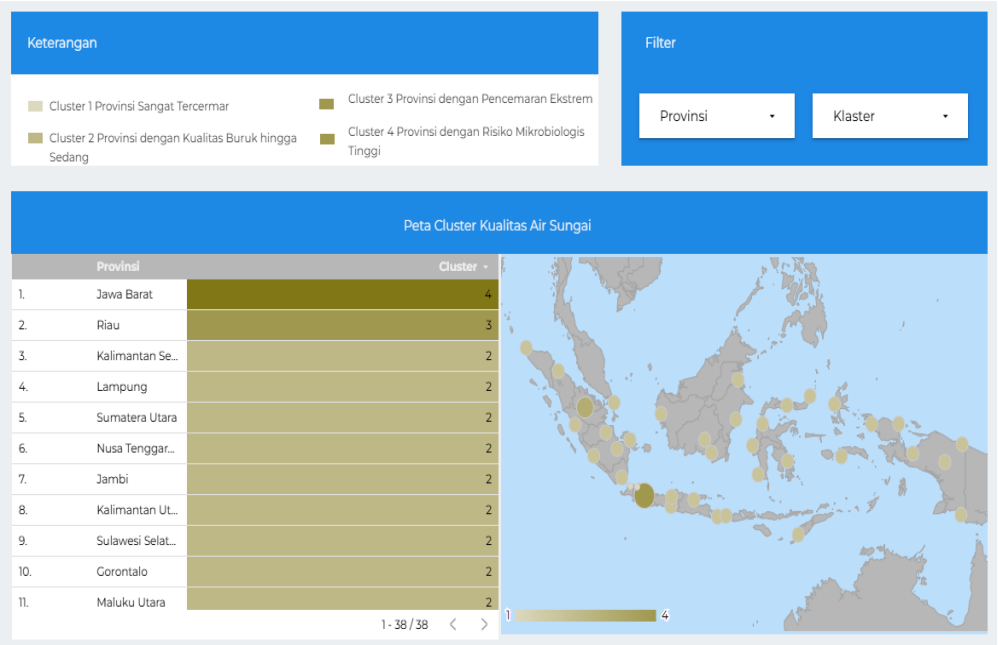
Berdasarkan hasil analisis clustering dan evaluasi parameter terhadap baku mutu, dapat disimpulkan bahwa tidak ada provinsi yang memiliki kualitas air sesuai dengan Kelas I. Hal ini mengindikasikan bahwa air sungai di seluruh provinsi Indonesia tidak dapat langsung digunakan sebagai bahan baku air minum tanpa pengolahan. Mayoritas provinsi termasuk dalam Kelas III dan IV, sehingga air hanya layak untuk pertanian, perikanan, atau pengairan. Provinsi dalam Klaster 1, 3, dan 4 terutama DKI Jakarta, Banten, Riau, dan Jawa Barat memerlukan penanganan serius melalui kebijakan pengendalian pencemaran, pembangunan instalasi pengolahan limbah cair, dan penegakan hukum lingkungan. Untuk provinsi dalam Klaster 2, diperlukan strategi pencegahan pencemaran melalui pendekatan pengelolaan berbasis DAS (daerah aliran sungai), pengawasan kegiatan industri, serta edukasi masyarakat tentang pentingnya menjaga kualitas air.

Pemerintah daerah perlu mengembangkan sistem pengelolaan limbah domestik dan industri yang terpadu serta meningkatkan kesadaran publik. Di sisi lain, pemerintah pusat disarankan untuk memperluas jaringan pemantauan kualitas air secara real-time agar dapat mendeteksi perubahan kondisi dengan cepat dan mengambil langkah mitigasi yang tepat. Dengan pendekatan berbasis data dan kebijakan yang responsif, perbaikan kualitas air nasional dapat tercapai secara bertahap dan berkelanjutan. Validasi menggunakan data resmi Statistik Lingkungan Hidup BPS 2024, konsistensi hasil *clustering* dengan status mutu sungai nasional.

Selain analisis historis, sistem pemantauan multiparameter berbasis sensor dapat memperkuat deteksi dini pencemaran karena mampu merekam parameter fisik-kimia dan indikator mikrobiologis secara berkelanjutan [18].

1. DKI Jakarta dan Banten (Klaster 1) sejalan dengan kondisi Sungai Ciliwung dan Cisadane yang tercatat "cemar ringan - sedang", namun data numerik menunjukkan pencemaran serius.
2. Riau (Klaster 3) tercatat "cemar ringan", tetapi parameter TSS dan COD mendukung kategori pencemaran ekstrem.
3. Provinsi Klaster 2 (Sumatera Utara, Sumatera Selatan, Jawa Tengah, Jawa Timur, NTB) konsisten dengan status "memenuhi baku mutu-cemar ringan".
4. Jawa Barat (Klaster 4) sesuai dengan kondisi Sungai Ciliwung dan Citarum yang tercemar mikrobiologis.

Validasi ini menegaskan bahwa analisis numerik melalui clustering memberikan representasi lebih presisi dibanding klasifikasi deskriptif semata. Visualisasi hasil segmentasi dilakukan untuk memperkuat interpretasi analisis *clustering* yang telah diuraikan sebelumnya. Visualisasi ini berperan penting dalam menyampaikan informasi secara komprehensif dan komunikatif kepada pemangku kebijakan maupun masyarakat umum, serta menjadi dasar pengambilan keputusan yang berbasis data.



Gambar 7. Peta *Cluster* Kualitas Air Sungai

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 4, Juni 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i4.9824

Salah satu visualisasi utama yang digunakan adalah peta interaktif yang menampilkan pembagian klaster kualitas air sungai di 38 provinsi Indonesia berdasarkan hasil pemodelan *Hierarchical Clustering*, sebagaimana terlihat pada Gambar. 6. Warna-warna pada peta digunakan untuk membedakan keempat klaster, yakni:
1. Klaster 1 (provinsi sangat tercemar),
2. Klaster 2 (provinsi dengan kualitas buruk hingga sedang),
3. Klaster 3 (provinsi dengan pencemaran ekstrem),
4. Klaster 4 (provinsi dengan risiko mikrobiologis tinggi).

Selanjutnya, visualisasi dalam bentuk grafik donat menampilkan distribusi jumlah provinsi dalam setiap klister. Hasil visual menunjukkan bahwa sebanyak 89,5% provinsi tergolong dalam Klaster 2, menandakan bahwa sebagian besar wilayah memiliki kualitas air dalam kategori buruk hingga sedang.

Rata-rata Parameter per Cluster

Cluster_HC	1	3	4	2
Rata-Rata pH	7.09	6.09	7.46	7.5
Rata-Rata Suhu (°C)	28	28.4	26.97	28.52
Rata-Rata TDS (mg/L)	114.82	124.7	390.99	302.77
Rata-Rata TSS (mg/L)	32.11	1,026.76	27.11	81.35
Rata-Rata BOD (mg/L)	8.45	2.8	4.67	3.47
Rata-Rata COD (mg/L)	42.97	137.04	17	19.56
Rata-Rata DO (mg/L)	3.64	4.57	6.08	5.95
Rata-Rata NH_3 (mg/L)	0.14	0.29	0.14	0.07
Rata-Rata NO_3 (mg/L)	0.98	1.05	2.78	1.34
Rata-Rata SO_4 (mg/L)	10.6	2.28	10.6	11.45
Rata-Rata Fecal Coliform (jml/100 ml)	16,907	353.52	189,460	6,977.38
Rata-Rata Total Coliform (jml/100 ml)	291,782.43	1,035.37	157,391.22	14,235.09
IKA 2023	49.85	50.84	46.87	55.98

1 - 4 / 4

Gambar 8. Rata-rata Parameter per *Cluster*

Selain itu, grafik rata-rata parameter kualitas air per klaster disusun untuk memperlihatkan perbedaan karakteristik antar klaster secara kuantitatif, sebagaimana terlihat pada Gambar. 7. Parameter-parameter seperti Total Suspended Solids (TSS), Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), serta Total Coliform menunjukkan nilai yang jauh melampaui ambang batas baku mutu di beberapa klaster. Misalnya, Klaster 3 yang hanya terdiri dari Provinsi Riau memiliki nilai TSS sebesar 1.026,76 mg/L dan COD sebesar 137,04 mg/L, yang menandakan beban pencemaran organik dan sedimen yang sangat tinggi.

Visualisasi nilai Indeks Kualitas Air (IKA) per provinsi juga digunakan untuk menghubungkan hasil segmentasi dengan status mutu air secara keseluruhan. Tidak ditemukan provinsi yang memiliki nilai IKA > 70, yang berarti tidak ada wilayah yang memenuhi klasifikasi Kelas I (layak sebagai bahan baku air minum tanpa pengolahan lanjut). Visualisasi ini secara keseluruhan memberikan manfaat strategis dalam mendukung pengelolaan kualitas air sungai, di antaranya:
1. Menyediakan gambaran spasial dan kuantitatif atas kondisi mutu air yang dapat dimanfaatkan oleh pengambil kebijakan,
2. Mengidentifikasi wilayah-wilayah yang memerlukan intervensi segera,
3. Memfasilitasi komunikasi data kepada publik secara lebih efektif,
4. Menjadi dasar dalam perencanaan kebijakan lingkungan hidup berbasis data.

Visualisasi lengkap hasil segmentasi dapat diakses melalui tautan Looker Studio: https://lookerstudio.google.com/reporting/50ab3713-f2fd-47bc-81ca-8c4bc8262a82

4. KESIMPULAN

Penelitian ini berhasil melakukan segmentasi kualitas air sungai di 38 provinsi di Indonesia menggunakan integrasi data parameter fisik, kimia, dan mikrobiologis di ibu kota provinsi serta Indeks Kualitas Air (IKA) tahun 2023. Melalui kerangka kerja CRISP-DM, proses pengelompokan yang diawali dengan pra-pemrosesan data termasuk penanganan nilai hilang, pertimbangan outlier, dan normalisasi menggunakan Robust Scaler mampu memetakan karakteristik pencemaran air secara komprehensif. Berdasarkan evaluasi metrik internal, model Hierarchical Clustering dengan konfigurasi empat klaster terbukti menunjukkan performa yang lebih unggul dibandingkan K-Means. Keunggulan tersebut ditunjukkan oleh nilai Silhouette Score sebesar 0,5702 (vs K-Means: 0,5528), Davies-Bouldin Index (DBI) sebesar 0,3744 (vs K-Means: 0,5703), dan Dunn Index sebesar 0,6168 (vs K-Means: 0,3380). Selain memberikan metrik matematis yang lebih solid, Hierarchical Clustering juga menghasilkan

struktur dendrogram yang mampu memperlihatkan pemisahan antar-klaster secara tegas sekaligus mempermudah interpretasi hubungan spasial antarwilayah. Oleh karena itu, model ini ditetapkan sebagai konfigurasi paling optimal dalam segmentasi kualitas air nasional.Hasil segmentasi makro ini berhasil membagi wilayah Indonesia ke dalam empat klaster risiko kualitas air yang selaras dengan regulasi mutu air nasional Peraturan Pemerintah Nomor 22 Tahun 2021. Temuan utama penelitian mengonfirmasi kondisi lingkungan yang kritis: tidak ada satu pun provinsi di Indonesia yang memenuhi syarat mutu air Kelas I (layak konsumsi tanpa pengolahan), dengan mayoritas wilayah berada pada kategori Kelas III dan IV (tercemar sedang hingga berat). Hasil pemodelan data-driven ini telah divisualisasikan ke dalam dashboard interaktif Looker Studio untuk mempermudah diseminasi informasi dan mendukung pengambilan kebijakan berbasis bukti.Meskipun memberikan kontribusi penting bagi tata kelola sumber daya air, penelitian ini memiliki keterbatasan mendasar pada aspek data. Ukuran sampel yang dianalisis sangat terbatas, yakni hanya mencakup 38 entri data yang merepresentasikan kondisi sungai di tingkat ibu kota provinsi. Selain itu, model ini memiliki ketergantungan yang tinggi terhadap teknik imputasi spasial akibat banyaknya nilai hilang (missing values) pada beberapa parameter krusial, seperti variabel amonia (NH3) dan sulfat (SO4) yang kehilangan hingga 21 nilai dari total entri. Adanya manipulasi data melalui imputasi iteratif ini berpotensi menyembunyikan variabilitas lokal yang ekstrem di lapangan. Penelitian selanjutnya disarankan untuk memperluas cakupan titik pemantauan di tingkat kabupaten/kota serta menggunakan data pemantauan langsung secara real-time untuk meminimalkan bias imputasi.

REFERENCES

[1] World Health Organization, "Drinking-water," World Health Organization. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/drinking-water.
[2] A. Afandi, "Dampak Laju Pertumbuhan Penduduk terhadap Alam dan Lingkungan," Kompasiana. [Online]. Available: https://www.kompasiana.com/azizafandi6697/649ad694e1a1672c18036a12/dampak-laju-pertumbuhan-penduduk-terhadap-alam-dan-lingkungan.
[3] Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia, "Laporan Status Lingkungan Hidup Indonesia 2022," KLHK RI, Jakarta, 2022.
[4] I. P. Rachmawati, E. Riani, and A. Riadi, "Status mutu air dan beban pencemaran Sungai Krukut, DKI Jakarta," *J. Nat. Resour. Environ. Manag.*, vol. 10, no. 2, pp. 220–233, 2020.
[5] L. Purwandari, "Laporan Kinerja Direktorat Pengendalian Pencemaran Udara Direktorat Jenderal Pengendalian Pencemaran dan Kerusakan Lingkungan Tahun 2022," 2023. [Online]. Available: https://tanamanpangan.pertanian.go.id/assets/front/uploads/document/LAKIN DJTP 2022_UPDATE ATAP (2).pdf
[6] A. N. Azizah, T. Widiharih, and A. R. Hakim, "Kernel K-Means Clustering Untuk Pengelompokan Sungai di Kota Semarang Berdasarkan Faktor Pencemaran Air," *J. Gaussian*, vol. 11, no. 2, pp. 228–236, 2022.
[7] B. Warsito, S. Sumiyati, H. Yasin, and H. Faridah, "Evaluation of river water quality by using hierarchical clustering analysis," in *IOP Conference Series: Earth and Environmental Science*, 2021, pp. 1–7.
[8] M. N. Ahmed, S. Shahid, and A. El-Shafie, "Machine learning methods for better water quality prediction," *J. Hydrol.*, vol. 578, p. 124084, 2019, doi: 10.1016/j.jhydrol.2019.124084.
[9] H. Chen and M. Franklin, "Spatio-Temporal Modeling of Surface Water Quality Distribution in California (1956-2023)," *arXiv preprint arXiv:2311.12736*, 2023.
[10] F. Sulianta, *Buku Dasar Data Mining from A to Z*. Feri Sulianta, 2023. [Online]. Available: https://www.researchgate.net/publication/382274667_Basic_Data_Mining_From_A_to_Z
[11] I. P. Aldiansah and M. Akrom, "Effect of Virtual Sample Generation in Predicting Corrosion Inhibition Efficiency on Pyridazine," *J. Appl. Informatics Comput.*, vol. 9, no. 2, pp. 382–389, 2025.
[12] N. K. Zuhal, "Study Comparison K-Means Clustering Dengan Algoritma Hierarchical Clustering: AHC, K-Means Clustering, Study Comparison," in *Seminar Nasional Teknologi & Sains*, 2022, pp. 200–205.
[13] A. R. Damayanti and A. W. Wijayanto, "Comparison of hierarchical and non-hierarchical methods in clustering cities in Java Island using the human development index indicators year 2018," *Eig. Math. J.*, vol. 4, no. 1, pp. 8–17, 2021.
[14] I. Indra, N. Nur, M. Iqram, and N. Inayah, "Perbandingan K-Means dan Hierarchical Clustering dalam Pengelompokan Daerah Beresiko Stunting," *INOVTEK Polbeng - Seri Informatika*, vol. 8, no. 2, pp. 356–367, 2023.
[15] I. Yahya, G. N. A. Wibawa, and L. Laome, "Penggunaan korelasi cophenetic untuk pemilihan metode cluster berhierarki pada mengelompokkan kabupaten/kota berdasarkan jenis penyakit di Provinsi Sulawesi Tenggara tahun 2020," in *Seminar Nasional Sains dan Terapan VI*, 2022, pp. 1–16.
[16] I. R. Drl, Y. H. Chrisnanto, and F. R. Umbara, "Analisis cluster pada kelompok masyarakat yang rentan terhadap paparan Covid-19 menggunakan metode K-Means clustering dan visualiasi dengan SIG," *Informatics Digit. Expert*, vol. 4, no. 2, pp. 61–69, 2022.
[17] H. Malikhatin, A. Rusgiyono, and I. M. Di Asih, "Penerapan K-Modes Clustering dengan Validasi Dunn Index Pada Pengelompokan Karakteristik Calon TKI Menggunakan R-GUI," *J. Gaussian*, vol. 10, no. 3, pp. 359–366, 2021.
[18] D. B. Krklješ, "Multiparameter Water Quality Monitoring System for Continuous Monitoring of Fresh Waters," *arXiv preprint arXiv:2307.11630*, 2023.