Identification of Bengawan Solo River Water Quality Patterns Using K-Means Clustering Based on Physicochemical and Environmental Parameters

Identifikasi Pola Kualitas Air Sungai Bengawan Solo Menggunakan Klasterisasi K-Means Berdasarkan Parameter Fisik-Kimia dan Lingkungan

Widya Cholid Wahyudin [1*], Tole Sutikno[2], Rusydi Uma[3]

[1)] *Department of Informatics, Faculty of Industrial Technology, Ahmad Dahlan University, Indonesia*
[2)] *Department of Electrical Engineering, Faculty of Industrial Technology, Ahmad Dahlan University, Indonesia*
[3)] *Department of Computer Science, Faculty of Science and Technology, Universitas Muhammadiyah Kudus, Indonesia*

[*]Coresponding author.
E-mail addresses: widyacholidwahyudin@umkudus.ac.id

Abstract. *River water quality needs to be monitored continuously because changes in physicochemical and environmental parameters may indicate early changes in aquatic conditions. This study aims to identify water quality patterns in the Bengawan Solo River using K-Means clustering based on physicochemical and environmental parameters. The dataset consists of 1,753 field observations with attributes including temperature, pH, electrical conductivity, total dissolved solids, water color, odor, and weather condition. The research stages include feature selection, data preprocessing, categorical encoding, Z-score standardization, K-Means clustering, and cluster number evaluation. The number of clusters was tested from K=2 to K=5. Cluster quality was evaluated using Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, and Inertia. After data cleaning, 1,751 observations were used in the clustering process. The evaluation results show that K=2 is the best cluster number, with a Silhouette Score of 0.187638 and a Calinski-Harabasz Score of 456.873808. The clustering results formed two main patterns, namely Cluster 0 with 840 observations or 47.97% and Cluster 1 with 911 observations or 52.03%. Based on average parameter characteristics, Cluster 0 has higher electrical conductivity and TDS values than Cluster 1; therefore, it is interpreted as a higher water quality risk pattern. These results indicate that K-Means can identify initial water quality patterns in an unlabeled Bengawan Solo River dataset.*

Keywords: Bengawan Solo, clustering, K-Means, TDS, water quality

Abstrak. *Kualitas air sungai perlu dipantau secara berkelanjutan karena perubahan parameter fisik-kimia dan lingkungan dapat menjadi indikator awal perubahan kondisi perairan. Penelitian ini bertujuan mengidentifikasi pola kualitas air Sungai Bengawan Solo menggunakan metode K-Means clustering berdasarkan parameter fisik-kimia dan lingkungan. Dataset yang digunakan terdiri atas 1.753 data lapangan dengan atribut suhu, pH, daya hantar listrik, total dissolved solids, warna air, bau air, dan kondisi cuaca. Tahapan penelitian meliputi seleksi fitur, preprocessing data, categorical encoding, standardisasi Z-score, klasterisasi K-Means, dan evaluasi jumlah klaster. Jumlah klaster diuji pada K=2, K=3, K=4, dan K=5. Evaluasi dilakukan menggunakan Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, dan Inertia. Setelah proses pembersihan data, jumlah data yang digunakan dalam klasterisasi menjadi 1.751 data. Hasil evaluasi menunjukkan bahwa K=2 merupakan jumlah klaster terbaik dengan Silhouette Score sebesar 0.187638 dan Calinski-Harabasz Score sebesar 456.873808. Hasil klasterisasi membentuk dua pola utama, yaitu Cluster 0 sebanyak 840 data atau 47.97% dan Cluster 1 sebanyak 911 data atau 52.03%. Berdasarkan karakteristik rata-rata parameter, Cluster 0 memiliki nilai DHL dan TDS lebih tinggi dibandingkan Cluster 1, sehingga diinterpretasikan sebagai pola risiko kualitas air lebih tinggi. Hasil ini menunjukkan bahwa K-Means dapat digunakan untuk mengidentifikasi pola awal kualitas air Sungai Bengawan Solo pada dataset tanpa label kelas independen.*

Kata kunci- Bengawan Solo, K-Means, kualitas air, klasterisasi, TDS

PENDAHULUAN

Sungai merupakan sumber daya air permukaan yang memiliki peran penting bagi kehidupan masyarakat, kegiatan pertanian, industri, dan keberlanjutan ekosistem. Kualitas air sungai dapat berubah akibat aktivitas antropogenik, limbah domestik, limbah industri, limpasan permukaan, perubahan tata guna lahan, serta kondisi lingkungan sekitar. Oleh karena itu, pemantauan kualitas air perlu dilakukan secara berkelanjutan agar perubahan parameter fisik-kimia dan lingkungan dapat diketahui lebih awal. Berbagai kajian terbaru menunjukkan bahwa machine learning

semakin banyak digunakan dalam pengelolaan sumber daya air karena mampu membantu analisis, prediksi, klasifikasi, dan interpretasi pola kualitas air pada data yang kompleks [1], [2], [3], [4]

Kualitas air umumnya dianalisis melalui beberapa parameter seperti suhu, pH, daya hantar listrik, total dissolved solids, dissolved oxygen, biological oxygen demand, chemical oxygen demand, serta parameter biologis lainnya. Namun, pada data lapangan, tidak semua parameter selalu tersedia secara lengkap. Selain itu, data kualitas air sering kali belum memiliki label kelas independen, seperti kategori baik, tercemar ringan, atau tercemar berat. Kondisi ini menyebabkan pendekatan klasifikasi supervised tidak selalu dapat langsung diterapkan, terutama ketika peneliti belum memiliki label kualitas air yang diperoleh dari uji laboratorium lengkap atau penilaian ahli [5], [6], [7], [8].

Pada kondisi dataset tanpa label kelas, pendekatan unsupervised learning dapat digunakan untuk mengidentifikasi pola awal dalam data. Salah satu metode yang banyak digunakan adalah K-Means clustering. Metode ini mengelompokkan data berdasarkan kemiripan karakteristik antarparameter, sehingga dapat membantu menemukan pola kualitas air tanpa harus membentuk label buatan. Pendekatan clustering juga relevan untuk pengelompokan sungai, analisis karakteristik kualitas air, dan pemodelan lingkungan yang bersifat heterogen [9], [10], [11], [12].

Penggunaan label buatan yang dibentuk dari skor parameter perlu dilakukan secara hati-hati. Apabila label dibuat dari parameter yang sama, kemudian parameter atau skor pembentuk label tersebut digunakan kembali dalam pemodelan, maka dapat terjadi data leakage. Data leakage dapat menyebabkan hasil model terlihat terlalu optimistis dan tidak mencerminkan kemampuan generalisasi pada data baru [13]. Oleh karena itu, penelitian ini tidak menggunakan label buatan sebagai target klasifikasi, tetapi menggunakan K-Means untuk mengidentifikasi pola kualitas air secara unsupervised.

Penelitian ini menggunakan K-Means clustering untuk mengidentifikasi pola kualitas air Sungai Bengawan Solo berdasarkan parameter suhu, pH, daya hantar listrik, total dissolved solids, warna air, bau air, dan kondisi cuaca. Sebelum proses klasterisasi, data melalui tahap preprocessing, categorical encoding, dan standardisasi Z-score. Jumlah klaster diuji pada K=2 sampai K=5, kemudian dievaluasi menggunakan Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, dan Inertia.

Penelitian sebelumnya menunjukkan bahwa pendekatan machine learning dapat digunakan untuk membantu proses analisis data dan pengambilan keputusan berbasis data. Wahyudin et al. [14] membandingkan beberapa model data mining dengan dukungan feature selection untuk meningkatkan performa prediksi dan mengurangi kompleksitas model. Pada konteks pemantauan lingkungan, Hernanda et al. [15] menunjukkan bahwa teknologi komputasi dan IoT dapat mendukung pemantauan kondisi lingkungan secara lebih sistematis. Berdasarkan landasan tersebut, penelitian ini menempatkan K-Means sebagai metode awal untuk mengidentifikasi pola kualitas air pada dataset lapangan yang belum memiliki label kelas independen.

Selain penelitian tersebut, Sutikno et al. [16] menunjukkan bahwa K-Means dapat digunakan secara kompetitif dalam strategi imputasi data ketika karakteristik atribut memiliki variasi tertentu, sehingga relevan dengan tahapan preprocessing dan pemilihan metode berbasis klaster pada penelitian ini. Umar et al. [17] menerapkan K-Means untuk pengelompokan data penyakit jantung, sedangkan Umar et al. [18] menggunakan K-Means dalam sistem identifikasi berbasis citra. Kedua penelitian tersebut memperlihatkan bahwa K-Means dapat dimanfaatkan untuk menemukan struktur kelompok pada data yang belum memiliki label kelas eksplisit.

Penelitian lain yang melibatkan Wahyudin et al. [19], [20], [21] juga menunjukkan bahwa pendekatan komputasi, data mining, dan Naive Bayes dapat digunakan untuk mendukung pengembangan sistem berbasis data serta pengambilan keputusan. Walaupun konteks penerapannya berbeda, pendekatan tersebut memperkuat landasan bahwa metode komputasi dapat dikembangkan secara bertahap, mulai dari pengolahan data, pembentukan pola, hingga pemodelan prediktif.

Dalam konteks kualitas air, berbagai penelitian menunjukkan bahwa Random Forest, ensemble learning, multi-class classification, dan explainable artificial intelligence dapat memperkuat analisis kualitas air serta membantu interpretasi parameter dominan [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]. Namun, penelitian-penelitian tersebut umumnya berangkat dari data berlabel atau model prediktif langsung. Oleh karena itu, penelitian ini menempatkan K-Means sebagai tahap awal untuk membentuk pola kualitas air pada dataset lapangan yang belum memiliki label kelas independen.

Kontribusi penelitian ini adalah mengidentifikasi pola kualitas air Sungai Bengawan Solo menggunakan K-Means clustering pada dataset tanpa label kelas independen. Hasil penelitian diharapkan dapat menjadi dasar untuk penelitian lanjutan, khususnya klasifikasi pola risiko kualitas air menggunakan model supervised learning dan pengembangan model explainable artificial intelligence pada tahap berikutnya.

METODE PENELITIAN

Dataset dan Preprocessing

Penelitian ini menggunakan pendekatan unsupervised learning dengan metode K-Means clustering untuk mengidentifikasi pola kualitas air Sungai Bengawan Solo. Dataset yang digunakan merupakan data kualitas air Sungai Bengawan Solo sebanyak 1.753 data dengan 10 atribut. Sepuluh atribut pada dataset awal terdiri atas `id_data`, `suhu_c`, `ph`, `dhl_us_cm`, `tds_mg_l`, `warna_encoded`,

bau_encoded, cuaca_encoded, tma_lokal_m, dan debit_manual_m3_detik. Namun, penelitian ini hanya menggunakan tujuh fitur utama dalam proses klasterisasi. Atribut id_data tidak digunakan karena hanya berfungsi sebagai identitas data, sedangkan tma_lokal_m dan debit_manual_m3_detik tidak digunakan dalam pemodelan utama karena ketersediaan datanya tidak lengkap. Pada penelitian ini, fitur yang digunakan untuk proses klasterisasi meliputi tujuh parameter utama, yaitu suhu air, pH, daya hantar listrik, total dissolved solids, warna air, bau air, dan kondisi cuaca.

Tahap preprocessing dilakukan untuk memastikan data siap diproses menggunakan algoritma K-Means. Pemeriksaan data menunjukkan terdapat 2 missing value pada parameter pH dan 2 missing value pada parameter TDS. Data yang memiliki nilai kosong pada fitur yang digunakan kemudian dihapus, sehingga jumlah data yang diproses dalam klasterisasi menjadi 1.751 data. Variabel kategorikal seperti warna air, bau air, dan kondisi cuaca direpresentasikan ke bentuk numerik melalui categorical encoding agar dapat diproses oleh algoritma *machine learning* ditunjukkan oleh Tabel 1.

Tabel 1. Fitur penelitian

No	Feature	Description
1	suhu_c	Suhu air dalam derajat Celsius
2	ph	Tingkat keasaman air
3	dhl_us_cm	Daya hantar listrik air
4	tds_mg_l	Jumlah zat padat terlarut
5	warna_encoded	Representasi numerik warna air
6	bau_encoded	Representasi numerik bau air
7	cuaca_encoded	Representasi numerik kondisi cuaca

Standardisasi dan K-Means Clustering

Sebelum proses klasterisasi, seluruh fitur distandardisasi menggunakan metode Z-score normalization. Standardisasi dilakukan karena K-Means merupakan algoritma berbasis jarak, sehingga perbedaan skala antarfitur dapat memengaruhi hasil klasterisasi. Parameter seperti TDS dan DHL memiliki rentang nilai lebih besar dibandingkan pH, sehingga tanpa standardisasi kedua parameter tersebut dapat mendominasi perhitungan jarak.

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

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2} \quad (2)$$

$$c_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i \quad (3)$$

$$J = \sum_{j=1}^{K} \sum_{i=1}^{m_j} \|x_i - c_j\|^2 \quad (4)$$

Pada Persamaan (1), z merupakan nilai hasil standardisasi, x adalah nilai asli data, μ adalah rata-rata fitur, dan σ adalah standar deviasi fitur. K-Means mengelompokkan data ke dalam beberapa klaster berdasarkan kemiripan karakteristik data. Jarak antara data dan centroid dihitung menggunakan Euclidean Distance seperti pada Persamaan (2), sedangkan pembaruan centroid ditunjukkan pada Persamaan (3). Fungsi objektif K-Means adalah meminimalkan jumlah kuadrat jarak antara data dan centroid seperti pada Persamaan (4).

Pada penelitian ini, jumlah klaster diuji pada K=2, K=3, K=4, dan K=5 untuk menentukan jumlah klaster yang paling sesuai dengan karakteristik data kualitas air Sungai Bengawan Solo. Evaluasi jumlah klaster dilakukan menggunakan Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, dan Inertia. Jumlah klaster terbaik ditentukan berdasarkan kombinasi nilai Silhouette Score tertinggi, Davies-Bouldin Index terendah, dan Calinski-Harabasz Score tertinggi.

Gambar 1. Alur penelitian identifikasi pola kualitas air menggunakan K-Means clustering

HASIL DAN PEMBAHASAN

Hasil Preprocessing Data

Dataset yang digunakan dalam penelitian ini terdiri atas 1.753 data dengan 10 atribut. Setelah dilakukan seleksi fitur, penelitian ini menggunakan tujuh parameter utama, yaitu suhu air, pH, daya hantar listrik, total dissolved solids, warna air, bau air, dan kondisi cuaca. Hasil pemeriksaan data menunjukkan bahwa terdapat 2 missing value pada parameter pH dan 2 missing value pada parameter TDS. Data yang memiliki nilai kosong pada fitur yang digunakan kemudian dihapus agar proses klasterisasi tidak dipengaruhi oleh nilai yang tidak lengkap. Setelah proses pembersihan data, jumlah data yang digunakan dalam analisis menjadi 1.751 data.

Tahap preprocessing juga mencakup penggunaan data hasil encoding pada variabel kategorikal, yaitu warna_encoded, bau_encoded, dan cuaca_encoded.

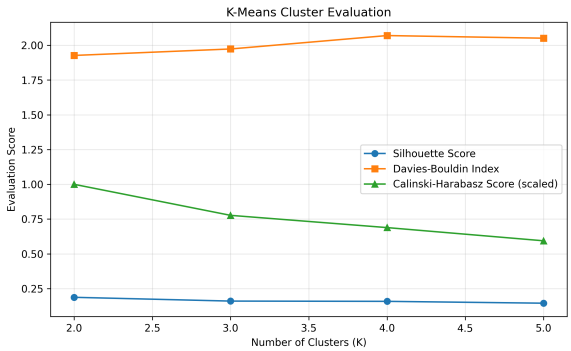
Encoding dilakukan agar data kategorikal dapat diproses oleh algoritma K-Means. Selanjutnya, seluruh fitur distandardisasi menggunakan Z-score karena K-Means merupakan algoritma berbasis jarak yang sensitif terhadap perbedaan skala antarfitur.

Evaluasi Jumlah Klaster

Proses klasterisasi dilakukan dengan menguji beberapa jumlah klaster, yaitu K=2, K=3, K=4, dan K=5. Evaluasi jumlah klaster dilakukan menggunakan Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Score, dan Inertia. Hasil evaluasi ditunjukkan pada Tabel 2.

Tabel 2. Hasil evaluasi K-Means

K	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score	Inertia
2	0.187638	1.926465	456.873808	9718.367807
3	0.160465	1.972924	354.929149	8717.034670
4	0.158596	2.069388	314.704099	7956.944060
5	0.145326	2.050436	271.311654	7558.762937



Gambar 2. Evaluasi jumlah klaster K-Means

Berdasarkan Tabel 2, jumlah klaster K=2 memperoleh nilai Silhouette Score tertinggi, yaitu 0.187638. Nilai Davies-Bouldin Index pada K=2 juga merupakan yang terendah, yaitu 1.926465, sedangkan Calinski-Harabasz Score pada K=2 merupakan yang tertinggi, yaitu 456.873808. Hal ini menunjukkan bahwa pembagian data menjadi dua klaster memberikan struktur pengelompokan yang paling representatif dibandingkan jumlah klaster lainnya.

Nilai Inertia mengalami penurunan ketika jumlah klaster bertambah. Hal ini merupakan kondisi yang wajar pada K-Means karena semakin banyak jumlah klaster, jarak data terhadap centroid cenderung semakin kecil. Namun, penentuan jumlah klaster terbaik tidak hanya didasarkan pada Inertia, tetapi juga mempertimbangkan kualitas pemisahan klaster melalui Silhouette Score, Davies-Bouldin Index, dan Calinski-Harabasz Score. Oleh karena itu, penelitian ini memilih K=2 sebagai jumlah klaster terbaik.

Distribusi dan Karakteristik Klaster

Distribusi dan Karakteristik Klaster ditunjukkan oleh Tabel 3, sedangkan Distribusi data pada setiap klaster ditunjukkan oleh Gambar 3.

Tabel 3. Distribusi data hasil klasterisasi

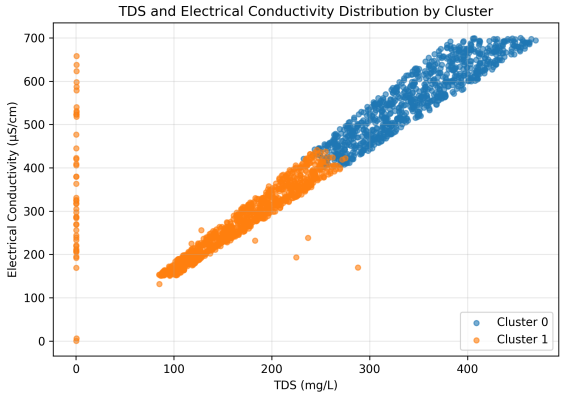
Kluster	Jumlah data	Presentase
Kluster 0	840	47.97%
Kluster 1	911	52.03%

Gambar 3. Distribusi data pada setiap klaster

Berdasarkan Tabel 3, Cluster 0 terdiri atas 840 data atau 47.97%, sedangkan Cluster 1 terdiri atas 911 data atau 52.03%. Distribusi ini menunjukkan bahwa data terbagi relatif seimbang ke dalam dua kelompok. Kondisi ini memperlihatkan bahwa hasil klasterisasi tidak didominasi oleh satu klaster tertentu, sehingga kedua kelompok dapat dianalisis lebih lanjut untuk memahami karakteristik masing-masing pola kualitas air.

Tabel 4. Karakteristik utama klaster

Kluster	Rata-rata DHL	Rata-rata TDS	Interpretasi
Kluster 0	558.7875	345.6404	Higher-risk pattern
Kluster 1	284.8330	164.5026	Lower-risk pattern



Gambar 4. Sebaran TDS dan daya hantar listrik berdasarkan klaster

Berdasarkan Tabel 4, Cluster 0 memiliki rata-rata daya hantar listrik sebesar 558.7875 dan rata-rata TDS sebesar 345.6404. Nilai tersebut lebih tinggi dibandingkan Cluster 1 yang memiliki rata-rata daya hantar listrik sebesar 284.8330 dan rata-rata TDS sebesar 164.5026. Perbedaan ini menunjukkan bahwa Cluster 0 memiliki kandungan zat terlarut dan kemampuan hantar listrik yang lebih tinggi dibandingkan Cluster 1.

Daya hantar listrik menggambarkan kemampuan air

dalam menghantarkan arus listrik yang berkaitan dengan kandungan ion terlarut, sedangkan TDS menunjukkan jumlah zat padat terlarut dalam air. Semakin tinggi nilai daya hantar listrik dan TDS, semakin besar indikasi adanya kandungan zat terlarut dalam air. Oleh karena itu, Cluster 0 diinterpretasikan sebagai pola risiko kualitas air lebih tinggi, sedangkan Cluster 1 diinterpretasikan sebagai pola risiko kualitas air lebih rendah. Interpretasi ini tidak dimaknai sebagai status pencemaran resmi, melainkan sebagai pola kualitas air berbasis klasterisasi.

Pembahasan

Hasil penelitian menunjukkan bahwa K-Means dapat digunakan untuk mengidentifikasi pola awal kualitas air Sungai Bengawan Solo pada dataset yang belum memiliki label kelas independen. Pendekatan ini sesuai dengan kondisi data yang belum memiliki kategori resmi seperti baik, tercemar ringan, atau tercemar berat. Dengan demikian, klasterisasi menjadi metode yang tepat untuk menemukan struktur pola secara alami tanpa memaksakan label buatan.

Pemilihan K=2 menunjukkan bahwa data kualitas air Sungai Bengawan Solo dalam penelitian ini cenderung membentuk dua pola utama. Pola pertama memiliki nilai TDS dan daya hantar listrik yang lebih tinggi, sedangkan pola kedua memiliki nilai TDS dan daya hantar listrik yang lebih rendah. Temuan ini menunjukkan bahwa kedua parameter tersebut menjadi pembeda utama dalam pembentukan pola kualitas air.

Nilai Silhouette Score yang relatif rendah menunjukkan bahwa pemisahan antar klaster belum terlalu kuat. Hal ini dapat terjadi karena kualitas air sungai merupakan fenomena lingkungan yang bersifat dinamis dan gradual. Perubahan parameter air tidak selalu terjadi secara tajam, tetapi dapat dipengaruhi oleh kondisi aliran, cuaca, aktivitas masyarakat, serta faktor lingkungan lain di sekitar sungai. Oleh karena itu, hasil klasterisasi perlu dipahami sebagai identifikasi pola awal, bukan sebagai penetapan status pencemaran resmi.

Secara metodologis, hasil penelitian ini dapat menjadi dasar untuk penelitian lanjutan. Klaster yang terbentuk pada Paper 1 dapat digunakan sebagai dasar pembentukan label pola risiko pada Paper 2. Dengan demikian, Paper 1 berperan sebagai tahap awal untuk mengidentifikasi pola kualitas air, sedangkan Paper 2 dapat dikembangkan untuk menguji konsistensi pola tersebut menggunakan metode klasifikasi supervised learning.

KESIMPULAN

Penelitian ini menunjukkan bahwa metode K-Means clustering dapat digunakan untuk mengidentifikasi pola kualitas air Sungai Bengawan Solo berdasarkan parameter fisik-kimia dan lingkungan. Dataset awal terdiri atas 1.753 data, kemudian setelah proses pembersihan missing value pada parameter pH dan TDS, data yang digunakan dalam proses klasterisasi berjumlah 1.751 data. Hasil evaluasi jumlah klaster menunjukkan bahwa K=2 merupakan jumlah klaster terbaik dibandingkan K=3, K=4, dan K=5, dengan Silhouette Score sebesar 0.187638, Davies-Bouldin Index sebesar 1.926465, dan Calinski-Harabasz Score sebesar 456.873808. Hasil klasterisasi membentuk dua pola utama, yaitu Cluster 0 sebanyak 840 data atau 47.97% dan Cluster 1 sebanyak 911 data atau 52.03%. Berdasarkan karakteristik rata-rata parameter, Cluster 0 memiliki nilai daya hantar listrik dan TDS lebih tinggi dibandingkan Cluster 1, sehingga diinterpretasikan sebagai pola risiko kualitas air lebih tinggi, sedangkan Cluster 1 diinterpretasikan sebagai pola risiko kualitas air lebih rendah. Temuan ini menunjukkan bahwa daya hantar listrik dan TDS berperan penting dalam membedakan pola kualitas air Sungai Bengawan Solo. Hasil penelitian ini tidak dimaknai sebagai status pencemaran resmi, tetapi sebagai pola awal kualitas air berbasis klasterisasi. Penelitian selanjutnya dapat menggunakan hasil klasterisasi ini sebagai dasar pembentukan label pola risiko untuk klasifikasi menggunakan metode supervised learning.

REFERENSI

[1] F. Ghobadi and D. Kang, "Application of Machine Learning in Water Resources Management: A Systematic Literature Review," *Water (Basel).*, vol. 15, no. 4, p. 620, 2023, doi: 10.3390/w15040620.

[2] A. Lokman, W. Z. W. Ismail, and N. A. A. Aziz, "A Review of Water Quality Forecasting and Classification Using Machine Learning Models and Statistical Analysis," *Water (Basel).*, vol. 17, no. 15, p. 2243, 2025, doi: 10.3390/w17152243.

[3] X. Yan, T. Zhang, W. Du, Q. Meng, X. Xu, and X. Zhao, "A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years," *J. Mar. Sci. Eng.*, vol. 12, no. 1, p. 159, 2024, doi: 10.3390/jmse12010159.

[4] P. Yuan *et al.*, "Optimizing water quality index using machine learning: A six-year comparative study in riverine and reservoir systems," *Sci. Rep.*, vol. 15, p. 33919, 2025, doi: 10.1038/s41598-025-10187-8.

[5] A. del Castillo, C. Yebra-Montes, M. Verduzco Garibay, J. de Anda, A. Garcia-Gonzalez, and M. S. Gradilla-Hernández, "Simple prediction of an ecosystem-specific water quality index and the water quality classification of a highly polluted river through supervised machine learning," *Water (Basel).*, vol. 14, no. 8, p. 1235, 2022, doi: 10.3390/w14081235.

[6] D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, "Using machine learning models for predicting the water quality index in the La Buong River, Vietnam," *Water (Basel).*, vol. 14, no. 10, p. 1552, 2022, doi: 10.3390/w14101552.

[7] A. Masood, M. Niazkar, M. Zakwan, and R. Piraei, "A machine learning-based framework for water quality index estimation in the Southern Bug River," *Water (Basel).*, vol. 15, no. 20, p. 3543, 2023, doi: 10.3390/w15203543.

[8] I. I. S. Shamsuddin, Z. Othman, and N. S. Sani, "Water

quality index classification based on machine learning: A case from the Langat River Basin model," *Water (Basel).*, vol. 14, no. 19, p. 2939, 2022, doi: 10.3390/w14192939.
[9] M. A. Novianta, S. Syafrudin, and B. Warsito, "K-Means clustering for grouping rivers in DIY based on water quality parameters," *JUITA: Jurnal Informatika*, vol. 11, no. 1, pp. 155–163, 2023, doi: 10.30595/juita.v11i1.16986.
[10] A. M. Abadi and others, "Determining river water quality in the Special Region of Yogyakarta using K-Means and Fuzzy C-Means algorithms," *TEM Journal*, 2025.
[11] Z. Zheng and others, "Spatiotemporal Prediction of Water Quality," *Environmental Pollution*, 2025.
[12] D. B. Laucelli, L. Enríquez, J. Saldarriaga, and O. Giustolisi, "Using symbolic machine learning to assess and model substance transport and decay in water distribution networks," *Sci. Rep.*, vol. 14, p. 3194, 2024, doi: 10.1038/s41598-024-53746-1.
[13] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
[14] W. C. Wahyudin, T. Sutikno, R. Umar, and A. Ridwan, "Comparison of Data Mining Model Performance in Heart Disease Detection with Feature Selection Application," *JOINCS (Journal of Informatics, Network, and Computer Science)*, vol. 8, no. 1, pp. 87–93, 2025.
[15] T. Hernanda, S. S. P. Nugroho, T. I. Izzati, F. K. Nisa, W. C. Wahyudin, and E. Nuriyatman, "IOT-BASED LEGAL POLICY IN CO_2 EMISSION SAFETY CONTROL TO SUPPORT GREEN TRANSPORTATION," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 9, no. 4, pp. 1434–1443, Oct. 2025, doi: 10.22437/jiituj.v9i4.48755.
[16] M. Muhammad, T. Sutikno, and I. Riadi, "A Comparative Study of K-Means and KNN Imputation for Handling Missing Data in Scholarship Applicant Datasets," *JUITA: Jurnal Informatika*, vol. 13, no. 3, 2025, doi: 10.30595/juita.v13i3.26502.
[17] J. Wala, H. Herman, and R. Umar, "Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 205–216, 2024, doi: 10.14421/jiska.2024.9.3.205-216.
[18] R. Umar, I. Riadi, and M. Miladiah, "Sistem Identifikasi Keaslian Uang Kertas Rupiah Menggunakan Metode K-Means Clustering," *Techno.Com*, vol. 17, no. 2, pp. 179–185, 2018, doi: 10.33633/tc.v17i2.1681.
[19] W. Cholid Wahyudin and S. P. Afrisia, "Design And Construction Of Shuff Photo Studio E-Booking Application Based On Responsive Web Rancang Bangun Aplikasi E-Booking Shuff Photo Studio Berbasis Web Responsif," 2024.
[20] W. C. Wahyudin, F. M. Hana, and A. Prihandono, "Prediksi Stunting Pada Balita Di Rumah Sakit Kota Semarang Menggunakan Naive Bayes," *Jurnal Ilmu Komputer dan Matemtika*, vol. 2019, pp. 32–36, 2023.
[21] A. Ridwan, T. Sutikno, I. Riyadi, and W. C. Wahyudin, "On-Time Student Graduation Prediction Modeling: A Comparative Analysis of Naive Bayes Algorithm and Other Data Mining Classifications," *JOINCS (Journal of Informatics, Network, and Computer Science)*, vol. 8, no. 2, 2025.
[22] J. Xu *et al.*, "An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies," *Water (Basel).*, vol. 13, no. 22, p. 3262, 2021, doi: 10.3390/w13223262.
[23] B. Schäfer *et al.*, "Machine learning approach towards explaining water quality dynamics in an urbanised river," *Sci. Rep.*, vol. 12, p. 12346, 2022, doi: 10.1038/s41598-022-16342-9.
[24] Y. B. Tran, L. F. Arias-Rodriguez, and J. Huang, "Predicting high-frequency nutrient dynamics in the Danube River with surrogate models using sensors and Random Forest," *Frontiers in Water*, vol. 4, p. 894548, 2022, doi: 10.3389/frwa.2022.894548.
[25] E. Dritsas and others, "Efficient Data-Driven Machine Learning Models for Water Quality Prediction," *Computers*, vol. 12, no. 2, p. 16, 2023, doi: 10.3390/computers12020016.
[26] M. M. Hassan *et al.*, "Efficient Prediction of Water Quality Index (WQI) Using Machine Learning Algorithms," *Human-Centric Intelligent Systems*, vol. 1, pp. 86–97, 2021.
[27] F. Firdiani, S. Mandala, Adiwijaya, and A. H. Abdullah, "WaQuPs: A ROS-Integrated Ensemble Learning Model for Precise Water Quality Prediction," *Applied Sciences*, vol. 14, no. 1, p. 262, 2024, doi: 10.3390/app14010262.
[28] L. Gao *et al.*, "Development and evaluation of a multi-class model for water quality assessment using machine learning," *Sci. Rep.*, vol. 15, p. 4785, 2025, doi: 10.1038/s41598-025-88799-5.
[29] V. Sangwan and R. Bhardwaj, "Machine Learning Framework for Predicting Water Quality Classification," *Water Pract. Technol.*, vol. 19, no. 11, pp. 4499–4521, 2024, doi: 10.2166/wpt.2024.259.
[30] A. Aldrees and others, "Evaluation of Water Quality Indexes with Novel Machine Learning and SHapley Additive ExPlanation (SHAP) Approaches," *Journal of Water Process Engineering*, vol. 59, p. 104789, 2024, doi: 10.1016/j.jwpe.2024.104789.
[31] H. A. Madni, M. Umer, and others, "Water-Quality Prediction Based on H2O AutoML and Explainable AI Techniques," *Water (Basel).*, vol. 15, no. 3, p. 475, 2023.
[32] M. K. Nallakaruppan, E. Gangadevi, M. L. Shri, B. Balusamy, S. Bhattacharya, and S. Selvarajan, "Reliable Water Quality Prediction and Parametric Analysis Using Explainable AI Models," *Sci. Rep.*, vol. 14, p. 7520, 2024, doi: 10.1038/s41598-024-56775-y.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relation- ships that could be construed as a potential conflict of interest.

Article History:
Received: 03 March 2026 | Accepted: 26 April 2026 | Published: 30 April 2026