https://jurnal.pustakagalerimandiri.co.id/index.php/pustakaai
DOI : https://doi.org/10.55382/jurnalpustakaai.v5i2.1114

JURNAL PUSTAKA AI
JURNAL PUSAT AKSES KAJIAN TEKNOLOGI ARTIFICIAL INTELLIGENCE

Vol. 5 No. 2 (2025) 166 – 176 E ISSN: 2809-4069

Klasifikasi Kualitas Air dengan *K-Means* dan *Decision Tree*

Anisa[1], Ahmad Turmudi Zy [2], Wahyu Hadikristanto[3]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Jl. Inspeksi Kalimalang Tegal Danas Arah Deltamas, Cibatu, Cikarang, Kabupaten Bekasi
[1]anisa478064@gmail.com [2]turmudi@pelitabangsa.ac.id [3]wahyu.hadikristanto@pelitabangsa.ac.id

Abstract

The problem addressed in this study started from the crucial need to ensure safe and consumable water quality which is often difficult to monitor in real time due to environmental fluctuations and limitations of manual systems. Therefore, this study aims to develop an accurate and automated water quality classification system by integrating the K-Means Clustering and Decision Tree methods. A dataset of 3,270 samples was collected through pH, Total Dissolved Solids (TDS), and temperature sensors, and processed using a data mining approach. The K-Means method was used to cluster unlabeled data while the Decision Tree algorithm was applied to predict water suitability based on the standards of Indonesia's Ministry of Health Regulation No. 2 of 2023. Evaluation results show a Silhouette Score of 0.8338, indicating good clustering quality and Decision Tree accuracy of 99.45% with an AUC value close to 1, confirming excellent classification performance. This study contributes to the development of an intelligent, data-driven IoT-based water quality monitoring system.

Keywords: AUC, K-Means, Decision Tree, Silhouette Score, Water Quality

Abstrak

Permasalahan dalam penelitian ini berangkat dari pentingnya menjaga kualitas air yang layak konsumsi, namun seringkali sulit untuk dipantau secara real-time karena fluktuasi lingkungan dan keterbatasan sistem manual. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan sistem klasifikasi kualitas air yang akurat dan otomatis dengan mengintegrasikan metode K-Means Clustering dan Decision Tree. Dataset sebanyak 3.270 sampel dikumpulkan melalui sensor pH, Total Dissolved Solids (TDS), dan suhu, yang kemudian diproses menggunakan pendekatan data mining. Metode K-Means digunakan untuk melakukan clustering terhadap data tanpa label, sedangkan Decision Tree digunakan untuk memprediksi kelayakan air berdasarkan standar Permenkes No. 2 Tahun 2023. Hasil evaluasi menunjukkan nilai Silhouette Score sebesar 0,8338 yang mencerminkan kualitas clustering yang baik, serta akurasi Decision Tree sebesar 99,45% dengan nilai AUC mendekati 1, yang menunjukkan performa klasifikasi yang sangat tinggi. Penelitian ini memberikan kontribusi dalam mendukung sistem pemantauan kualitas air berbasis IoT yang cerdas dan berbasis data.

Kata kunci: AUC, Decision Tree, Kualitas air, K-Means, Silhouette Score

1. Pendahuluan

Perkembangan teknologi informasi dan kecerdasan buatan telah membuka peluang besar dalam bidang pemantauan lingkungan, salah satunya dalam mengelola kualitas air.Air bersih yang layak konsumsi merupakan kebutuhan dasar masyarakat. Di Indonesia, parameter seperti pH, Total Dissolved Solids (TDS), dan suhu digunakan sebagai indikator utama kualitas air berdasarkan standar Permenkes No. 2 Tahun 2023. Namun, kondisi kualitas air di lapangan seringkali mengalami fluktuasi akibat faktor lingkungan maupun aktivitas manusia,

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

sehingga diperlukan sistem pemantauan yang mampu bekerja secara otomatis dan real-time. Di sinilah teknologi Internet of Things (IoT) dan kecerdasan buatan (Artificial Intelligence/AI) memainkan peran penting.[1]

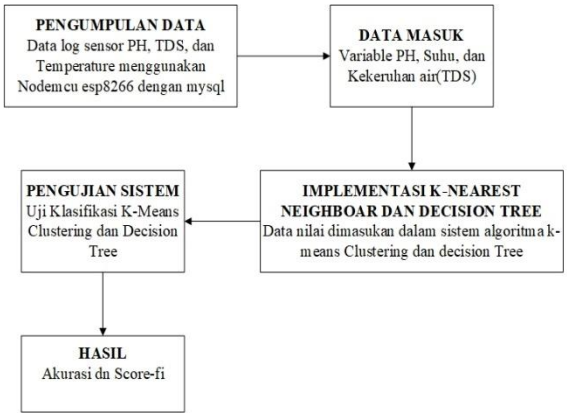
Berbagai penelitian sebelumnya telah mengkaji pendekatan klasifikasi kualitas air menggunakan metode berbasis machine learning. Hardiana Said,dkk. (2022) mengembangkan kombinasi K-Means dan Decision Tree yang menghasilkan akurasi tinggi dalam pengklasifikasian data air [2]. Nurmahaludin dan Gunawan (2019) menunjukkan bahwa K-Nearest Neighbor unggul dalam klasifikasi kualitas air, sementara K-Means lebih efektif dalam pengelompokan data [3]. Lidya Savitri dan Rahmat Nursalim (2023) menegaskan bahwa Decision Tree memiliki keunggulan dalam memetakan keputusan klasifikasi kualitas air dibandingkan SVM dan Logistic Regression [4].

Kendati demikian, belum banyak penelitian yang secara nyata menerapkan kombinasi metode unsupervised (K-Means) dan supervised (Decision Tree) secara terintegrasi dalam sistem klasifikasi kualitas air berbasis sensor IoT secara real-time. Hal ini menjadi celah penelitian (gap) yang penting untuk diisi.Berdasarkan hal tersebut, penelitian ini memiliki *novelty* dalam bentuk integrasi algoritma K-Means dan Decision Tree untuk klasifikasi kualitas air yang diimplementasikan pada sistem IoT secara real-time. Sistem ini dirancang menggunakan NodeMCU ESP8266 sebagai penghubung sensor ke database, yang memungkinkan pengambilan keputusan berbasis data yang cepat dan akurat. Evaluasi sistem dilakukan dengan pendekatan kuantitatif, termasuk pengukuran Silhouette Score, akurasi klasifikasi, serta analisis ROC dan AUC. [5].

Tujuan utama dari penelitian ini adalah untuk Mengimplementasikan K-Means dalam mengelompokkan kualitas air berdasarkan data sensor; Mengkaji akurasi Decision Tree dalam menentukan kelayakan konsumsi air; dan Mengevaluasi performa integrasi kedua algoritma dalam membangun sistem monitoring kualitas air berbasis IoT. Penelitian ini juga diarahkan untuk mendukung pengembangan sistem klasifikasi kualitas air yang lebih efisien, yang dapat diimplementasikan pada wilayah-wilayah dengan keterbatasan infrastruktur teknologi, serta menjadi referensi awal bagi penelitian lanjutan yang mengintegrasikan teknologi IoT dan kecerdasan buatan dalam pemantauan lingkungan secara real-time.

2. Metode Penelitian

Metode penelitian adalah serangkaian langkah atau proses yang disusun secara sistematis dan logis untuk melaksanakan penelitian demi mencapai tujuan yang telah ditetapkan[6]. Penelitian ini menggunakan metode kuantitatif dengan pendekatan data mining. Dataset berisi 3.270 sampel yang dikumpulkan dari sensor pH, TDS, dan suhu yang terhubung dengan NodeMCU ESP8266. Proses meliputi: preprocessing data, clustering menggunakan K-Means dengan validasi Silhouette Score, serta klasifikasi menggunakan Decision Tree dengan evaluasi menggunakan Confusion Matrix, ROC Curve, dan AUC yang berfokus pada penentuan kelayakan kualitas air dengan pendekatan berbasis data menggunakan metode K-Means dan Decision Tree. Proses-proses yang diterapkan dalam penelitian ini disajikan pada Gambar 1.



Gambar 1. Diagram Blok Metode Penelitian

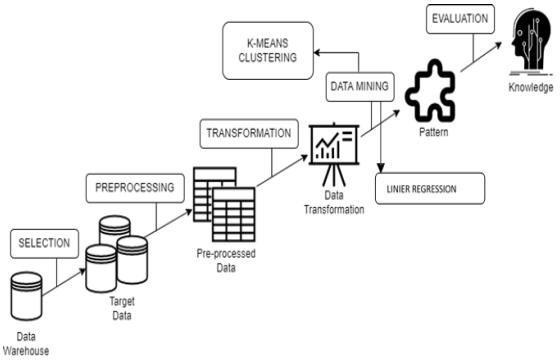
Dimulai dari proses seleksi data yang diambil dari database hasil pembacaan sensor pH, TDS, dan suhu secara real-time, dilakukan pengecekan dan validasi awal untuk memastikan kualitas data. Tahap praproses (preprocessing) mencakup seleksi atribut, normalisasi data menggunakan teknik standard scaling, serta transformasi data kualitatif menjadi numerik dengan label encoder. Setelah itu, data dipersiapkan untuk pemodelan dengan menggabungkan seluruh atribut relevan dalam satu dataset.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Selanjutnya, model K-Means digunakan untuk proses clustering dengan evaluasi menggunakan Silhouette Score guna menilai kualitas pembentukan klaster. Seperti yang diterapkan oleh [7] pemanfaatan K-Means dalam membentuk dua klaster berdasarkan atribut penilaian terbukti dapat memberikan hasil yang optimal dan dapat diadaptasi pada skenario klasifikasi berbasis parameter kualitas air.Model Decision Tree diterapkan untuk klasifikasi kelayakan air berdasarkan standar Permenkes. Seluruh hasil pemodelan dianalisis lebih lanjut untuk mengidentifikasi pola hubungan antar variabel dan kelompok klaster. Visualisasi hasil juga dilakukan untuk mendukung interpretasi.Selain itu, sebagai pendukung, analisis regresi linier digunakan untuk mengevaluasi hubungan antar variabel dan membantu memahami kontribusi setiap variabel terhadap kualitas air. Pendekatan ini memberikan landasan yang kuat untuk menyusun sistem pemantauan kualitas air yang objektif, terukur, dan siap diimplementasikan.

2.1. Tahap Implementasi

Pada tahapan implementasi penulis menggunakan metode Knowledge Discovery in Database[8], Berikut adalah gambar diagram proses pada Knowledge Discovery in Database :

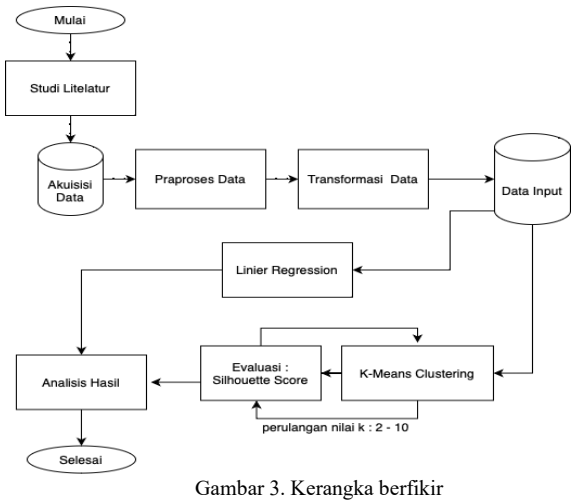


Gambar 2. Proses Knowledge Discovery in Database Process (KDD)

Knowledge Discovery in Database (KDD) merupakan penemuan pengetahuan dalam database. Untuk lebih spesifiknya KDD didefinisikan sebagai proses ekstraksi atau identifikasi pola, pengetahuan dan informasi potensial dari sekumpulan data yang besar. KDD ini cocok digunakan untuk mencari pola dan hubungan dari data yang besar untuk menghasilkan informasi baru[9].

2.2. Kerangka Berfikir

Penelitian ini menggunakan pendekatan berbasis data dalam menentukan kelayakan kualitas air dengan memanfaatkan metode K-Means untuk clustering dan Decision Tree untuk klasifikasi disajikan pada Gambar 3.



Gambar 3. Kerangka berfikir

Proses diawali dengan studi literatur dan akuisisi data dari sensor pH, TDS, dan suhu. Data yang diperoleh melalui tahap prapemrosesan dan transformasi sebelum dimasukkan ke sistem. Evaluasi dilakukan menggunakan metrik Silhouette Score untuk clustering dan akurasi untuk klasifikasi. Analisis terhadap hasil model menghasilkan informasi objektif yang mendukung sistem klasifikasi kualitas air secara sistematis.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

3. Hasil dan Pembahasan

Pengujian yang dilakukan dalam penelitian ini dilakukan secara pengambilan data yang sudah tersimpan pada database digunakan sebagai dataset yang strukturnya belum memiliki label dan tabel.Pengumpulan data dilakukan dengan mengekspor 3.270 record dari database sistem monitoring berbasis sensor pH, TDS, dan suhu ke dalam format CSV. Dataset yang diperoleh belum memiliki label kualitas air, dan berisi atribut seperti index, ID, pH, temperatur, TDS, dan timestamp. Data ini kemudian disiapkan untuk tahap preprocessing dan analisis lebih lanjut. Data dikumpulkan dalam format CSV dan divisualisasikan untuk memastikan tidak ada outlier ekstrem yang disajikan pada tabel 1.

3.1. Pre-procesing Data

Tabel 1.Tabel Dataset 5 Awal Dan 5 Akhir

index	id	ph	temperatur	tds	timestamp
0	1	7.2	25.3	300	11/12/2024 8:00
1	2	6.8	26.1	320	11/12/2024 9:00
2	3	7	25.8	310	11/12/2024 10:00
3	4	7.5	26.5	330	11/12/2024 11:00
4	5	7.1	27	295	11/12/2024 0:00
….	….	….	….	….	….
3265	3266	6.97	20.6	38	02/23/2025 16:37:38 AM
3266	3267	6.97	31	38	02/23/2025 16:37:43 AM
3267	3268	6.97	27.8	38	02/23/2025 16:37:48 AM
3268	3269	6.97	25.4	35	02/23/2025 16:37:53 AM
3269	3270	6.96	32.6	25	02/23/2025 16:37:58 AM

Data yang telah dikumpulkan melalui sensor kemudian diproses melalui tahap preprocessing, yang meliputi pembersihan data dari duplikasi dan outlier (data cleaning) yang disajikan pada gambar 4.

```
# Memuat ulang dataset dengan pemisah yang benar (;)
df = pd.read_csv(file_path, delimiter=";")

# Memilih hanya kolom yang diperlukan
df = df[["ph", "temperature", "tds"]]

# Mengonversi tipe data ke numerik
df = df.apply(pd.to_numeric, errors='coerce')

# Menampilkan dataset yang telah dibersihkan
df.head()
```

	ph	temperature	tds
0	7.2	25.3	300.0
1	6.8	26.1	320.0
2	7.0	25.8	310.0
3	7.5	26.5	330.0
4	7.1	27.0	295.0

Gambar 4. Proses Source code menghilangkan noise

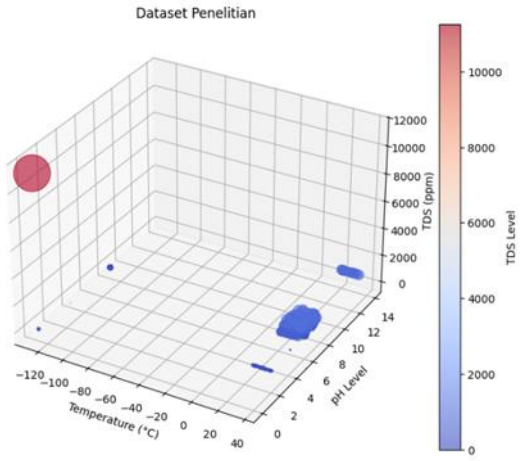
konversi data kategorikal menjadi numerik menggunakan label encoder (transformation), dan normalisasi nilai pH, suhu, serta TDS agar memiliki skala yang seragam. Karena tipe data ‘string’ tidak dapat digunakan dalam pemodelan Decision Tree dan K-Means, maka data kategori seperti status kualitas air akan diubah menjadi angka agar dapat diproses dalam pemodelan. Hasil dari transformasi data tersebut adalah sebagai berikut:

	ph	ph_category	ph_category_num	temperature	temperature_category	temperature_category_num	tds	tds_category	tds_category_num
0	7.2	Netral	1	25.3	Normal	1	300.0	Sedang	1
1	6.8	Netral	1	26.1	Normal	1	320.0	Sedang	1
2	7.0	Netral	1	25.8	Normal	1	310.0	Sedang	1
3	7.5	Netral	1	26.5	Normal	1	330.0	Sedang	1
4	7.1	Netral	1	27.0	Normal	1	295.0	Baik	0
5	7.1	Netral	1	26.9	Normal	1	659.0	Buruk	2
6	8.2	Netral	1	21.2	Normal	1	378.0	Sedang	1
7	7.3	Netral	1	26.5	Normal	1	487.0	Sedang	1
8	7.8	Netral	1	28.5	Normal	1	506.0	Sedang	1
9	6.7	Netral	1	28.4	Normal	1	634.0	Buruk	2

Gambar 5. Hasil dari transformasi

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

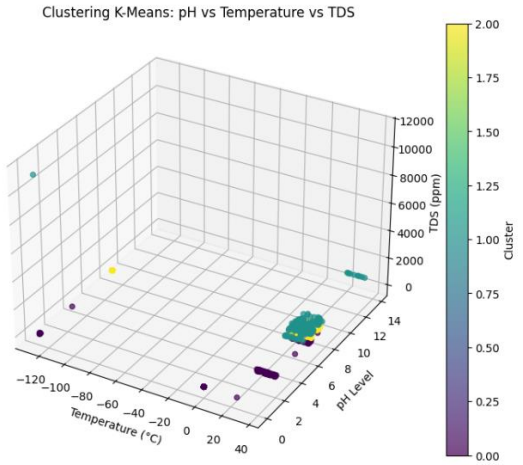
Proses ini dilakukan menggunakan Python di lingkungan Google Colab dan VSCode. Visualisasi awal menggunakan scatter plot digunakan untuk mengecek distribusi data sebelum dilakukan pemodelan. dilakukan untuk membersihkan dan menormalkan data, termasuk seleksi atribut dan konversi nilai kategorikal menjadi numerik[10].



Gambar 6. visualisasi dataset penelitian

3.2. Implementasi algoritma K-Means

Setelah melalui tahap pengecekan dan preprocessing, data kini siap untuk dianalisis lebih lanjut. Penelitian ini diawali dengan pemodelan dataset kualitas air menggunakan algoritma K-Means dalam bahasa pemrograman Python.Langkah pertama yang dilakukan adalah menentukan label untuk data dengan menerapkan metode K-Means clustering. Teknik ini digunakan untuk mengelompokkan data berdasarkan nilai pH, suhu, dan TDS, sehingga dapat mengidentifikasi pola kualitas air berdasarkan kelompok yang terbentuk.



Gambar 7. visualisasi clustering K-3

Setelah dilakukan pengecekan dan preprocessing data, proses penelitian ini dilanjutkan dengan penerapan algoritma K-Means untuk clustering data kualitas air. Data yang digunakan dalam proses clustering ini hanya mencakup parameter pH, suhu (temperature), dan TDS, dengan jumlah klaster k = 3.Jumlah klaster dibagi menjadi tiga untuk mengelompokkan data berdasarkan pola yang lebih terperinci dalam dataset.

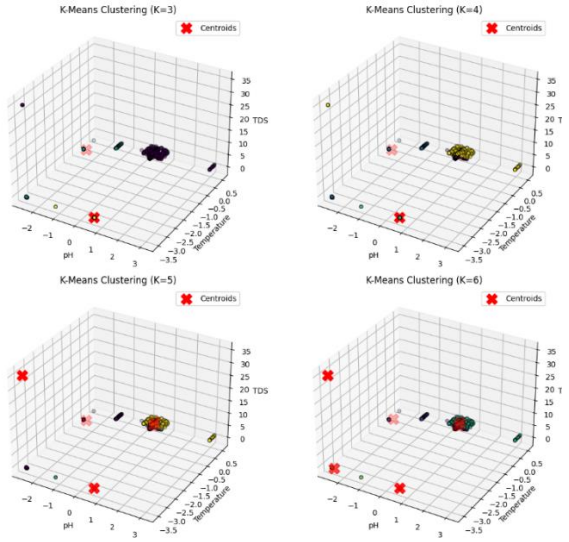
Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

```
[ ] from sklearn.preprocessing import StandardScaler

    # Normalisasi data menggunakan StandardScaler
    scaler = StandardScaler()
    df_scaled = scaler.fit_transform(df[["ph", "temperature", "tds"]])

    # Menampilkan beberapa data setelah normalisasi
    df_scaled[:5]
```

```
array([[0.41879692, 0.30324468, 0.10435762],
       [0.25281413, 0.32337042, 0.1683016 ],
       [0.33580553, 0.31582327, 0.13632961],
       [0.54328402, 0.33343329, 0.20027359],
       [0.37730122, 0.34601187, 0.08837162]])
```

Gambar 8. Centroid K-Means

Dengan menggunakan K = 3, dataset akan terbagi menjadi tiga kelompok utama, yang nantinya akan digunakan sebagai label dalam analisis kualitas air.Dari hasil clustering, diperoleh tiga kelompok dengan karakteristik sebagai berikut: Cluster_0 → Kualitas Baik (nilai rata-rata pH, suhu, dan TDS berada dalam rentang yang baik), Cluster_1 → Kualitas Sedang (nilai rata-rata pH, suhu, dan TDS berada dalam rentang menengah), Cluster_2 → Kualitas Buruk (nilai rata-rata pH, suhu, dan TDS menunjukkan kondisi yang kurang baik).



Gambar 9. clutering kmeans k-3 sampai K-6

Hasil dari proses clustering menunjukkan bahwa variabel clustering berhasil membagi dataset kualitas air ke dalam tiga kelompok utama berdasarkan parameter pH, suhu, dan TDS. Setiap klaster merepresentasikan pola tertentu dalam data, yang nantinya digunakan sebagai dasar dalam analisis lebih lanjut mengenai kualitas air.

```
cluster_results = {}

for K in K_values:
    kmeans = KMeans(n_clusters=K, random_state=42, n_init=10)
    df[f'Cluster_K{K}'] = kmeans.fit_predict(features_scaled)
    cluster_results[K] = df.groupby(f'Cluster_K{K}')[['ph', 'temperature', 'tds']].mean()

# Menampilkan rata-rata variabel numerik berdasarkan klaster untuk setiap K
cluster_results
```

```
{3:                ph  temperature         tds
Cluster_K3
0           7.069239    27.249116  319.746349
1           0.352993    -0.063858    4.691796
2           7.789770  -127.000000  185.115853,
4:                ph  temperature         tds
Cluster_K4
0           7.007711    27.301979  153.086490
1           0.352993    -0.063858    4.691796
2           7.789770  -127.000000  185.115853
3           7.199904    27.136855  673.673469,
5:                ph  temperature          tds
Cluster_K5
0           0.352993    -0.063858     4.691796
1           7.007323    27.296548   152.794567
2           7.789770  -127.000000   185.115853
3           0.000000  -127.000000  11257.000000
4           7.208897    27.333573   660.353717,
6:                ph  temperature          tds
Cluster_K6
0           7.007323    27.296548   152.794567
1           0.361458     3.405923     4.487472
2           7.810000  -127.000000   185.968241
3           7.208897    27.333573   660.353717
4           0.303077  -127.000000    11.307692
5           0.000000  -127.000000  11257.000000}
```

Gambar 10. Data cluster k3 sampai k6

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

Implementasi Agglomerative Clustering, sebuah metode dalam Hierarchical Clustering, menggunakan pustaka scikit-learn dalam Python. Proses ini diawali dengan mengimpor kelas AgglomerativeClustering dan menentukan jumlah klaster optimal (optimal_k). Model kemudian diterapkan pada dataset yang telah dinormalisasi (df_scaled), menghasilkan label klaster untuk setiap data. Setelah klastering selesai, hasilnya dievaluasi menggunakan Silhouette Score, sebuah metrik yang mengukur seberapa baik suatu data berada dalam klaster yang tepat. Nilai Silhouette Score berkisar antara -1 hingga 1, di mana nilai mendekati 1 menunjukkan klastering yang baik, sementara nilai mendekati 0 atau negatif menunjukkan kemungkinan tumpang tindih atau salah klasterisasi.

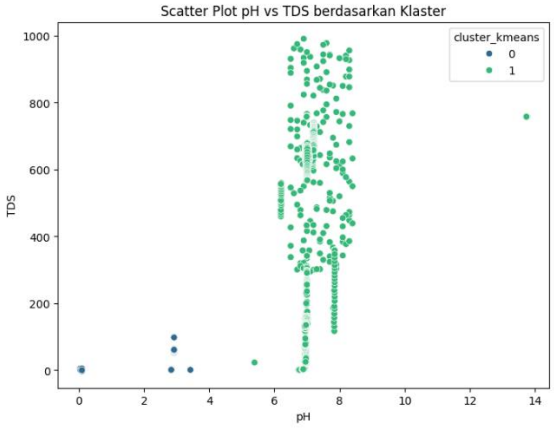
```
from sklearn.cluster import AgglomerativeClustering

agglo = AgglomerativeClustering(n_clusters=optimal_k)
labels = agglo.fit_predict(df_scaled)
silhouette_avg = silhouette_score(df_scaled, labels)

print(f"Silhouette Score dengan Agglomerative Clustering: {silhouette_avg}")
```

```
Silhouette Score dengan Agglomerative Clustering: 0.8315972958766603
```

Gambar 11. Hasil Clustering Silhouette Score

Dalam konteks penelitian tentang kualitas air, metode ini dapat digunakan untuk mengelompokkan data berdasarkan parameter pH, temperatur, dan TDS. Dengan menerapkan Agglomerative Clustering, dataset kualitas air dapat dibagi menjadi beberapa klaster, misalnya air layak konsumsi, tercemar ringan, dan tercemar berat. Evaluasi dengan Silhouette Score membantu memastikan bahwa jumlah klaster yang dipilih optimal, sehingga hasil klastering dapat digunakan untuk analisis lebih lanjut atau pengambilan keputusan dalam pemantauan kualitas air.

Nilai Silhouette Score yang dihasilkan adalah 0.83, yang menunjukkan bahwa klastering yang dilakukan memiliki kualitas yang sangat baik. Dengan nilai mendekati 1**,** ini berarti data dalam setiap klaster memiliki kemiripan tinggi, sementara jarak antar klaster cukup jauh. Hal ini menunjukkan bahwa pembagian klaster cukup jelas dan tidak ada banyak data yang berada di perbatasan antara klaster yang berbeda.untuk lebih jelas peneliti buatkan visualisasi dengan scatter plot untuk ph vs tds berdasarkan klister pada gambar 12.

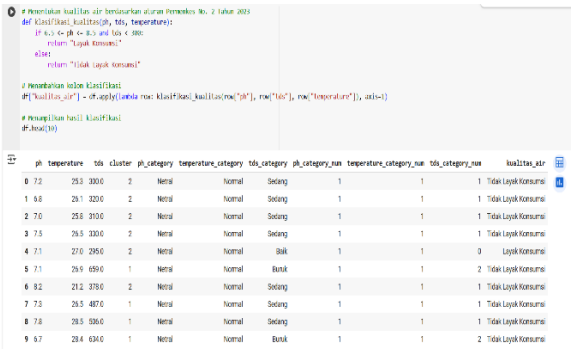


Gambar 12. scatter plot ph vs tds berdasarkan klaster

3.3. Implementasi Algoritma Decision Tree (Python)

Setelah dilakukan clustering data dengan metode K-Means, hasil model dari clustering tersebut digunakan sebagai label pada dataset kualitas air. Dataset yang sudah dilabelkan kemudian dilakukan pemodelan dengan menggunakan salah satu algoritma dalam metode klasifikasi, yaitu algoritma Decision Tree. Karena penelitian ini berfokus pada analisis apakah parameter pH, temperatur, dan TDS berpengaruh terhadap kualitas air, maka data yang digunakan dalam penelitian ini mencakup parameter pH, temperatur, dan TDS[11]. Model decision tree akan digunakan untuk mengklasifikasikan kualitas air berdasarkan parameter-parameter tersebut, sehingga dapat membantu dalam pengambilan keputusan mengenai kondisi air di lokasi penelitian.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

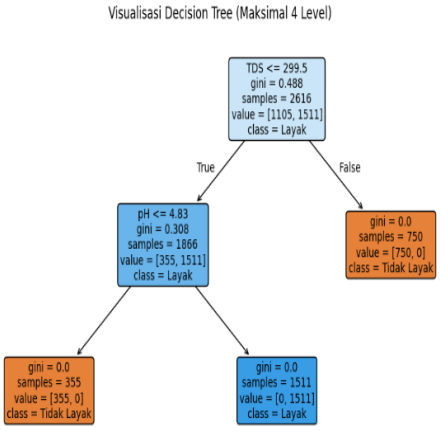


```
# Menentukan kualitas air berdasarkan aturan Permenkes No. 2 Tahun 2023
def klasifikasi_kualitas(ph, tds, temperature):
    if 6.5 <= ph <= 8.5 and tds < 300:
        return "Layak Konsumsi"
    else:
        return "Tidak Layak Konsumsi"

# Menambahkan kolom klasifikasi
df['kualitas_air'] = df.apply(lambda row: klasifikasi_kualitas(row['ph'], row['tds'], row['temperature']), axis=1)

# Menampilkan hasil klasifikasi
df.head(10)
```

	ph	temperature	tds	cluster	ph_category	temperature_category	tds_category	ph_category_num	temperature_category_num	tds_category_num	kualitas_air
0	7.2	25.3	330.0	2	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
1	6.8	29.1	320.0	2	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
2	7.0	25.8	310.0	2	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
3	7.5	26.5	330.0	2	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
4	7.1	27.0	295.0	2	Netral	Normal	Baik	1	1	0	Layak Konsumsi
5	7.1	26.9	659.0	1	Netral	Normal	Buruk	1	1	2	Tidak Layak Konsumsi
6	8.2	21.2	378.0	2	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
7	7.3	26.5	487.0	1	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
8	7.8	28.5	536.0	1	Netral	Normal	Sedang	1	1	1	Tidak Layak Konsumsi
9	6.7	28.4	634.0	1	Netral	Normal	Buruk	1	1	2	Tidak Layak Konsumsi

Gambar 13. Dataset Untuk Proses Decision Tree

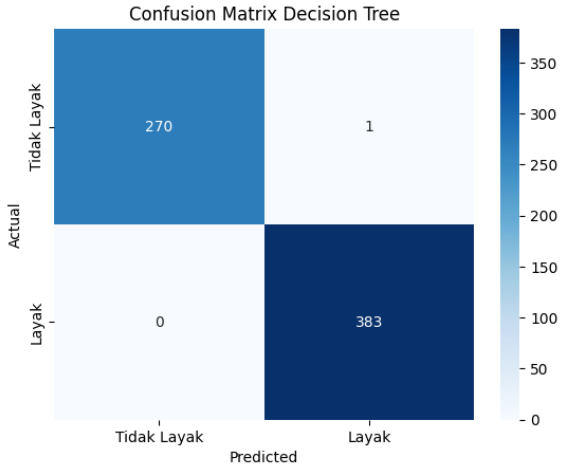
Pemodelan kualitas air menggunakan Decision Tree dilakukan dengan membagi dataset yang berisi nilai pH, suhu, dan TDS serta label kualitas air ("Baik", "Cukup", "Buruk") menjadi data pelatihan dan pengujian dengan rasio 8:2[12]. Model dilatih menggunakan kriteria entropy dan max_depth=3 untuk membatasi kedalaman pohon keputusan. Setelah model dilatih, hasil prediksi diuji pada data pengujian dan akurasinya dihitung. Pohon keputusan yang dihasilkan dapat diekspor untuk memahami bagaimana keputusan diambil berdasarkan parameter pH, suhu, dan TDS dalam menentukan kualitas air.



Gambar 14. scatter plot ph vs tds berdasarkan klaster

Gambar 4. 1 Diagram Pohon Decision Tree

untuk mengetahui bisa dengan Confusion Matrix digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi (y_pred) dan nilai aktual (y_test). Matriks ini dihitung menggunakan confusion_matrix dan ditampilkan sebagai heatmap menggunakan sns.heatmap, dengan label "Tidak Layak" dan "Layak" pada sumbu x dan y. Peta panas ini menunjukkan jumlah prediksi yang benar dan salah untuk kedua kelas, memberikan gambaran tentang akurasi model dalam mengklasifikasikan data.berikut gambar



Gambar 15. Confusion Matrix Decision Tree

3.4. Evaluasi

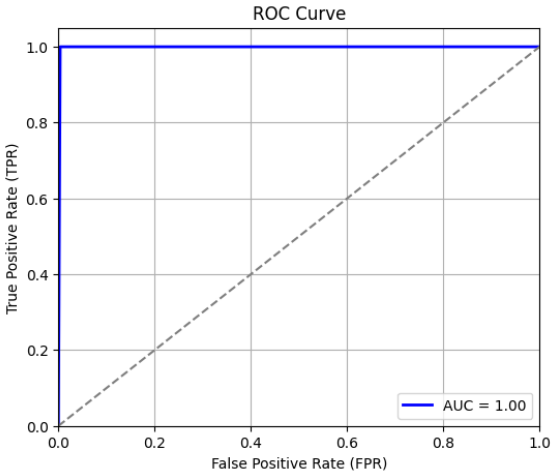
Pada tahap ini dilakukan uji kelayakan apakah algoritma decision tree bisa digunakan sebagai algoritma untuk memprediksi pada penelitian ini. Untuk mendapatkan jawaban tersebut, maka diperlukan beberapa fungsi dalam klasifikasi. Dimana pada pemodelan decision tree ini akan dilakukan prediksi terlebih dahulu, kemudian akan dilihat confusion matrix, classification report, nilai akurasi, nilai AUC dan juga kurva AUC[13]. Semua proses yang dilakukan menggunakan bahasa pemrograman python. Maka hasil dari pengujian ini adalah sebagai berikut:

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Mengonversi label ke bentuk numerik (Layak Konsumsi = 1, Tidak Layak Konsumsi = 0)
df["kualitas_air_label"] = df["kualitas_air"].map({"Layak Konsumsi": 1, "Tidak Layak Konsumsi": 0})

# Memisahkan fitur dan label
X = df[["ph", "temperature", "tds"]]
y = df["kualitas_air_label"]

# Membagi data menjadi training (80%) dan testing (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Membuat model Decision Tree
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)

# Memprediksi pada data uji
y_pred = dt_model.predict(X_test)

# Menghitung akurasi model
accuracy = accuracy_score(y_test, y_pred)
accuracy
```

```
0.9984709480122325
```

Gambar 16. Hasil Akurasi Pada Decision Tree

akurasi model Decision Tree mencapai 99%**,** itu menunjukkan bahwa model memiliki kinerja yang sangat baik dalam mengklasifikasikan data pengujian dengan benar. Namun, akurasi saja tidak selalu memberikan gambaran yang lengkap tentang kinerja model, terutama dalam kasus dataset yang tidak seimbang atau jika model cenderung mengklasifikasikan satu kelas lebih sering daripada kelas lainnya. Oleh karena itu, meskipun akurasi tinggi, penting untuk juga mengevaluasi metrik lain seperti confusion matrix**,** ROC curve**,** dan AUC score untuk memahami lebih dalam seberapa baik model dalam menangani kedua kelas (kelas "Layak" dan "Tidak Layak"). Dalam beberapa kasus, meskipun akurasi tinggi, model masih melakukan kesalahan signifikan dalam mengklasifikasikan kelas minoritas, yang bisa mengurangi keandalan keputusan yang dihasilkan.



Gambar 17. Diagram ROC Curva

untuk menghitung dan menampilkan ROC Curve (Receiver Operating Characteristic) serta AUC Score (Area Under the Curve) sebagai evaluasi kinerja model klasifikasi. predict_proba digunakan untuk mendapatkan probabilitas prediksi kelas positif ("Layak Konsumsi"), kemudian roc_curve menghitung False Positive Rate (FPR) dan True Positive Rate (TPR)[14]. AUC dihitung menggunakan fungsi auc, yang mengukur area di bawah kurva ROC. Grafik ROC menggambarkan trade-off antara FPR dan TPR, dengan AUC sebagai indikator kinerja model: semakin mendekati 1, semakin baik model dalam membedakan antara kelas[15].

3.5. Pembahasan

Dalam analisis kualitas air menggunakan algoritma Decision Tree dan K-Means Clustering**,** didapatkan hasil bahwa kedua metode ini menunjukkan performa yang sangat baik dalam melakukan klasifikasi dan pengelompokan kualitas air.Decision Tree berhasil mencapai akurasi 99%**,** yang menunjukkan bahwa model ini sangat efektif dalam memprediksi kelayakan kualitas air berdasarkan parameter yang digunakan. Hal ini menunjukkan bahwa karakteristik data yang digunakan memiliki pola yang kuat dan dapat diklasifikasikan dengan baik oleh model**.**K-Means Clustering dengan nilai Silhouette 8.3 mengindikasikan bahwa pemisahan antar klaster yang terbentuk cukup jelas, sehingga pengelompokan data kualitas air dapat dilakukan dengan akurasi yang tinggi.

4. Kesimpulan

Berdasarkan hasil dan pembahasan, dapat disimpulkan bahwa penelitian ini berhasil menjawab permasalahan pada latar belakang, yaitu perlunya sistem klasifikasi kualitas air berbasis data yang akurat dan real-time. Dengan memanfaatkan kombinasi algoritma K-Means untuk proses pengelompokan dan Decision Tree untuk klasifikasi, sistem mampu mengolah data sensor pH, suhu, dan TDS secara otomatis melalui perangkat NodeMCU ESP8266. Hasil clustering menunjukkan nilai Silhouette Score sebesar 0,8338, yang mengindikasikan kualitas klaster yang sangat baik, sementara akurasi klasifikasi oleh Decision Tree mencapai 99,45%, serta didukung oleh nilai AUC yang mendekati 1. Keberhasilan ini menunjukkan bahwa integrasi kedua algoritma tersebut mampu menghasilkan sistem monitoring kualitas air yang efektif, akurat, dan dapat diimplementasikan secara real-time pada berbagai kondisi lingkungan. Dengan demikian, penelitian ini memberikan kontribusi nyata dalam pengembangan teknologi pemantauan lingkungan berbasis IoT dan machine learning yang dapat diandalkan.

Daftar Rujukan

[1] Kementerian Kesehatan Republik Indonesia. (2023). Peraturan Menteri Kesehatan Nomor 2 Tahun 2023 tentang Persyaratan Kualitas Air Minum. Diakses dari https://peraturan.go.id
[2] Said, H., Matondang, N. H., & Irmanda, H. N. (2022). Sistem Prediksi Kualitas Air Yang Dapat Dikonsumsi dengan Menerapkan Algoritma K-Nearest Neighbor. Jurnal Teknologi Informasi, 14(2), 77–84.
[3] Savitri, L., & Nursalim, R. (2023). Komparasi Decision Tree, SVM, dan Logistic Regression dalam Klasifikasi Kualitas Air. Jurnal Teknik Informatika dan Sistem Informasi, 9(1), 33–40. https://doi.org/10.21067/smartics.v9i1.8113
[4] Hartanti, D., & Pradana, A. I. (2023). Komparasi Metode Decision Tree, Logistic Regression, SVM, dan ANN dalam Klasifikasi Kualitas Air. SMARTICS Journal, 9(1), 1–6. https://doi.org/10.21067/smartics.v9i1.8113
[5] Saputra, A. W., Purnamasari, A. I., & Ali, I. (2024). Implementasi Algoritma Naïve Bayes untuk Memprediksi Kualitas Air yang Dapat Dikonsumsi. Jurnal Teknologi dan Sistem Komputer, 12(1), 45–52.
[6] Sari, D. P., Buana, W., & Sari, M. F. M. (2025).Implementasi Data Mining pada Penjualan Barang dengan Teknik K-Means. Jurnal Pustaka AI, 5(1), 106–112.
[7] https://doi.org/10.55382/jurnalpustakaai.v5i1.955Kristanto, B., Zy, A. T., & Fatchan, M. (2023). Analisis penentuan karyawan tetap dengan algoritma K-Means dan Davies Bouldin Index. Bulletin of Information Technology (BIT), 4(1), 112–120. https://doi.org/10.47065/bit.v3i1.521

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025

[8] A. Tangkelayuk and E. Mailoa, "Klasifikasi Kualitas Air Menggunakan Metode KNN, Naïve Bayes Dan Decision Tree," vol. 9, no. 2, pp. 1109–1119, 2022, [Online]. Available: http://jurnal.mdp.ac.id
[9] D. Hartanti and A. I. Pradana, "Komparasi Metode Decission Tree, Logistic Regression, SVM, dan ANN Dalam Klasifikasi Kualitas Air," SMARTICS Journal, vol. 9, no. 1, pp. 1–6, Apr. 2023, doi: 10.21067/smartics.v9i1.8113.
[10] M. B. Abdul Majid, Y. M. Cani, and U. Enri, "Penerapan Algoritma K-Means dan Decision Tree Dalam Analisis Prestasi Siswa Sekolah Menengah Kejuruan," Jurnal Sistem Komputer dan Informatika (JSON), vol. 4, no. 2, p. 355, Dec. 2022, doi: 10.30865/json.v4i2.5299.
[11] M. C. Untoro and M. Muttaqin, "Detection Malaria Base Microscope Digital Image with Convolutional Neural Network," SinkrOn, vol. 7, no. 3, pp. 935–943, Jul. 2022, doi: 10.33395/sinkron.v7i3.11488.
[12] A. N. Vidyanti et al., "Symptom-based scoring technique by machine learning to predict COVID-19: a validation study," BMC Infect Dis, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12879-023-08846-0.
[13] A. W. Saputra, A. I. Purnamasari, and I. Ali, "IMPLEMENTASI ALGORITMA NAÏVE BAYES UNTUK MEMPREDIKSI KUALITAS AIR YANG DAPAT DI KONSUMSI," 2024.
[14] Sangaji, D., & Sutabri, T. (2025).Analisis XGBoost dan Random Forest untuk Prediksi Curah Hujan dalam Mendukung Mitigasi Karhutla. Jurnal Pustaka AI, 5(1), 13–18.
[15] https://doi.org/10.55382/jurnalpustakaai.v5i1.905V. Pramaningsih, R. Yuliawati, S. Sukisman, H. Hansen, R. Suhelmi, and A. Daramusseng, "Indek Kualitas Air dan Dampak terhadap Kesehatan Masyarakat Sekitar Sungai Karang Mumus, Samarinda," Jurnal Kesehatan Lingkungan Indonesia, vol. 22, no. 3, pp. 313–319, Oct. 2023, doi: 10.14710/jkli.22.3.313-319.

Submitted : 20-07-2025 | Reviewed : 25-08-2025 | Accepted : 02-08-2025