JDMIS: Journal of Data Mining and Information Systems
Vol. 3, No. 2, 2025, p. 86~93
ISSN: ISSN: 2986-3473 | p-ISSN 2986-5271
DOI: 10.54259/jdmis.v3i2.4513

Penerapan Algoritma K-Nearest Neighbors untuk Klasifikasi Kualitas Air Minum

Jansen[1], Cariven Tanova[2], Dariel[3], Marciano[4] Ade Maulana[5]
[1,2,3,4,5]Sistem Informasi (Kampus Kota Medan), Universitas Pelita Harapan, Indonesia
Email: [1]03081230024@student.uph.edu,, [2]03081230027@student.uph.edu, [3]03081230013@student.uph.edu,
[4]03081230020@student.uph.edu, [5]ade.maulana@lecturer.uph.edu

ABSTRAK

Penelitian ini bertujuan untuk mengklasifikasikan kelayakan air minum berdasarkan parameter fisik dan kimia menggunakan algoritma *K-Nearest Neighbors* (KNN). Data diambil dari platform Kaggle, terdiri dari 100.000 sampel air dengan sembilan atribut utama, termasuk pH, kekerasan, TDS, sulfat, chloramines, konduktivitas, karbon organik, trihalomethanes, dan kekeruhan. Label target adalah *potability*, yang menunjukkan apakah air layak dikonsumsi (1) atau tidak (0). Tahapan prapengolahan mencakup normalisasi dan pembagian data menjadi data latih dan uji. Model KNN dibangun dengan mengevaluasi berbagai nilai *K* untuk mendapatkan performa optimal. Hasil evaluasi menggunakan confusion matrix menunjukkan bahwa model mampu mencapai akurasi sebesar 78%. Pada kelas air layak, diperoleh precision sebesar 72%, recall 91%, dan F1-score 81%. Sementara itu, untuk air tidak layak, precision mencapai 88%, recall 65%, dan F1-score 75%. Meskipun model menunjukkan kecenderungan salah mengklasifikasikan air tidak layak sebagai layak, secara keseluruhan performanya cukup baik. Hasil ini menunjukkan bahwa KNN dapat digunakan sebagai pendekatan klasifikasi yang efektif dan berpotensi diterapkan dalam sistem pemantauan kualitas air secara otomatis.

Kata Kunci: Kualitas Air, Kelayakan Air Minum, K-Nearest Neighbors, Klasifikasi, Machine Learning

ABSTRACT

This study aims to classify drinking water potability based on physical and chemical parameters using the K-Nearest Neighbors *(KNN) algorithm. The dataset, sourced from the Kaggle platform, contains 100,000 water samples with nine key attributes, including pH, hardness, total dissolved solids (TDS), sulfate, chloramines, conductivity, organic carbon, trihalomethanes, and turbidity. The target label is* potability*, indicating whether the water is safe (1) or unsafe (0) for consumption. The preprocessing steps included normalization and splitting the data into training and testing sets. The KNN model was trained by experimenting with various K values to achieve optimal performance. Evaluation using a confusion matrix showed that the model achieved an accuracy of 78%. For the potable class, the model reached a precision of 72%, recall of 91%, and F1-score of 81%. For the non-potable class, it achieved a precision of 88%, recall of 65%, and F1-score of 75%. Although the model tends to misclassify unsafe water as safe, overall performance is promising. These findings suggest that the KNN algorithm can serve as an effective classification approach and has potential for application in automated water quality monitoring systems.*

Keywords: *Water Quality, Drinking Water Potability, K-Nearest Neighbors, Classification, Machine Learning*

Penulis Korespondensi:

Ade Maulana
Email: ade.maulana@lecturer.uph.edu

Article Info

Diterima: 27 Mei 2025
Direvisi: 28 Mei 2025
Disetujui: 28 Mei 2025

1. PENDAHULUAN

Air adalah kebutuhan utama bagi semua makhluk hidup [1]. Salah satu sumber air yang tersedia di sekitar kehidupan manusia adalah air sungai yang mengalir [2] . Ketersediaan air minum yang aman dan layak konsumsi merupakan salah satu indikator penting dalam pencapaian tujuan pembangunan berkelanjutan (SDGs), khususnya poin keenam yang menargetkan akses universal terhadap air bersih pada tahun 2030. Menurut standar air minum Indonesia yang ditetapkan oleh PP No. 82 Tahun 2001 dan KepMen No. 907 Tahun 2002, air bersih yang digunakan setiap hari harus berkualitas baik untuk dikonsumsi [3]. Namun, hingga saat ini masih terdapat sekitar dua miliar penduduk dunia yang menggunakan sumber air minum terkontaminasi [4]. Air yang tidak layak konsumsi dapat membawa berbagai risiko kesehatan serius, termasuk penyakit diare, kolera, disentri, dan infeksi parasit lainnya, yang menyumbang angka kematian cukup tinggi di berbagai negara berkembang [5]. Menurut laporan WHO, sekitar 829.000 orang meninggal setiap tahunnya akibat konsumsi air yang tercemar, sanitasi yang buruk, serta kurangnya kebersihan [6]. Di samping itu, kualitas air juga dipengaruhi oleh berbagai faktor fisik dan kimia, seperti tingkat pH, kandungan logam berat, serta parameter mikrobiologis yang tidak selalu dapat dideteksi secara visual [7]. Air disebut sebagai senyawa kompleks karena mengandung berbagai zat dan mineral. Namun, tidak semua kandungan tersebut dapat diserap oleh tubuh manusia. Air juga mudah tercemar oleh zat dan bakteri berbahaya akibat pencemaran sumber air atau lingkungan sekitarnya. Oleh karena itu, diperlukan pengawasan dan pengolahan yang ketat agar kualitas air tetap terjaga sesuai standar dan aman untuk dikonsumsi [8]. Pemantauan kualitas air secara rutin dan akurat menjadi tantangan tersendiri, terutama di wilayah dengan keterbatasan laboratorium dan teknologi pengujian [9].Meskipun diketahui presentasi 97% air yang hadir di bumi, hanya ada sekitar 27% yang secara layak untuk dilakukan pengonsumsian atau memenuhi kebutuhan konsumsi manusia [10]. Di Indonesia sendiri, data dari Badan Pusat Statistik menunjukkan bahwa pada tahun 2022, hanya sekitar 73,9% rumah tangga yang memiliki akses terhadap air minum layak, dan angka ini lebih rendah lagi di daerah perdesaan [11]. Kondisi ini memperkuat urgensi terhadap pengembangan sistem yang mampu memprediksi kelayakan air secara cepat dan efisien sebagai bagian dari upaya preventif dalam perlindungan kesehatan masyarakat.

Pengujian kualitas air dengan metode konvensional seringkali memakan waktu dan biaya yang tinggi, sehingga dibutuhkan pendekatan yang lebih efisien dan didukung oleh pemanfaatan data [12]. Salah satu pendekatan yang semakin banyak dimanfaatkan dalam membantu proses klasifikasi data lingkungan, termasuk dalam penilaian kualitas air, adalah *machine learning.* Metode ini memungkinkan komputer untuk mengenali pola dari sejumlah data historis dan menggunakannya untuk membuat prediksi terhadap data baru. Dalam studi ini, penulis menggunakan algoritma *K-Nearest Neighbor* (KNN) sebagai teknik klasifikasi guna memprediksi kelayakan air minum berdasarkan parameter kualitas air seperti pH, kandungan logam berat, dan senyawa kimia lainnya. KNN merupakan algoritma yang sederhana namun efektif, karena bekerja dengan cara menghitung kedekatan antara data uji dan sejumlah data pelatihan yang diketahui kelasnya, lalu mengklasifikasikan data uji berdasarkan label mayoritas dari tetangganya [13][14]. Sifat algoritma ini yang fleksibel serta tidak membutuhkan asumsi distribusi data menjadikannya cocok digunakan dalam permasalahan klasifikasi kualitas air yang melibatkan banyak variabel dan kemungkinan *outlier*.

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma KNN mampu memberikan performa yang baik dalam mengklasifikasikan kualitas air. Penelitian oleh Jadhav dan Channe berhasil mencapai akurasi sebesar 90,14% dalam memprediksi kelayakan air menggunakan metode KNN pada *dataset* kualitas air dari *Central Pollution Control Board* India[15]. Studi lainnya oleh Patel dan Thakkar juga menunjukkan bahwa KNN dapat digunakan secara efektif dalam sistem monitoring kualitas air berbasis sensor, dengan tingkat akurasi mencapai 85,23% [16]. Sementara itu, penelitian oleh Shukla et al. menyimpulkan bahwa KNN menjadi salah satu algoritma yang kompetitif dibanding metode lain seperti *Decision Tree* dan SVM dalam mengklasifikasikan air bersih dan tercemar, dengan akurasi 87,5% saat digunakan pada *dataset* yang diambil dari sumber sungai [17]. Penelitian-penelitian ini mengindikasikan bahwa KNN merupakan algoritma yang layak untuk digunakan dalam sistem prediksi kualitas air minum berbasis data.

Penelitian ini bertujuan untuk mengembangkan sebuah model klasifikasi kualitas air yang bersifat sederhana namun tetap akurat, dengan memanfaatkan tiga parameter utama yaitu pH, suhu, dan tingkat kekeruhan sebagai variabel masukan. Ketiga parameter ini dipilih karena dapat diperoleh dengan mudah melalui perangkat sensor dan merupakan indikator penting dalam menentukan kelayakan air. Diharapkan, hasil dari penelitian ini dapat memberikan kontribusi dalam upaya deteksi dini kualitas air, sehingga masyarakat maupun instansi terkait dapat melakukan tindakan cepat untuk mencegah dampak kesehatan dan lingkungan yang lebih luas.

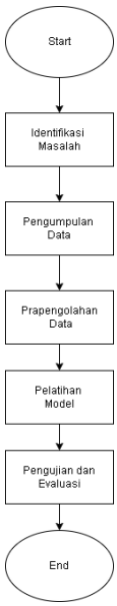
2. METODE PENELITIAN

Pada penelitian ini, proses klasifikasi kualitas air dilakukan dengan pendekatan algoritma *machine learning K-Nearest Neighbor* (KNN). Secara umum, tahapan penelitian ini terdiri dari lima langkah utama, yakni identifikasi masalah, pengumpulan data, prapengolahan data, pelatihan model, serta pengujian dan evaluasi. Kelima tahapan tersebut digambarkan dalam diagram alur proses sebagai dasar metodologi yang digunakan dalam penelitian ini.

2.1. Identifikasi Masalah

Pada tahap ini, dirumuskan permasalahan terkait meningkatnya kebutuhan akan sistem pemantauan kualitas air yang efisien dan akurat, khususnya dalam kaitannya dengan lingkungan dan kesehatan masyarakat. Penilaian kualitas air secara manual memerlukan waktu, tenaga, serta sumber daya yang besar, dan cenderung rawan terhadap kesalahan subjektif. Untuk itu,

dibutuhkan suatu metode klasifikasi otomatis yang mampu mengelompokkan kualitas air berdasarkan parameter fisik dan kimia tertentu. Pendekatan berbasis *machine learning* ini diharapkan dapat meningkatkan efektivitas pemantauan serta mendukung pengambilan keputusan yang lebih tepat dalam pengelolaan sumber daya air.



Gambar 1 Metode Penelitian

2.2. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, yang menyediakan *dataset* terkait kualitas air. Dataset tersebut terdiri dari 100.000 entri data yang masing-masing merepresentasikan sampel air dengan sembilan atribut utama, yaitu *pH, hardness, total dissolved solids (TDS), chloramines, sulfate, conductivity, organic carbon, trihalomethanes,* dan *turbidity*. Label target dari *dataset* ini adalah potability, yang mengindikasikan apakah air tersebut layak untuk dikonsumsi (1) atau tidak layak konsumsi (0). *Dataset* ini dipilih karena memenuhi beberapa kriteria penting, antara lain: (1) jumlah data yang cukup untuk mendukung pelatihan dan evaluasi model secara representatif; (2) atribut-atribut yang digunakan mencerminkan parameter kimia dan fisik air yang umum dipakai dalam standar penilaian kualitas air; serta (3) format data yang telah terstruktur dan siap untuk dilakukan prapengolahan dan klasifikasi lebih lanjut.

2.3. Prapengolahan Data

Sebelum membangun model klasifikasi kualitas air, dilakukan serangkaian proses prapengolahan data untuk memastikan data yang digunakan bersih, lengkap, dan dalam format yang sesuai untuk algoritma KNN. Berikut adalah tahapan prapengolahan data yang dilakukan dalam penelitian ini:

Tabel 1 Tahap Prapengolahan Data

No	Tahapan Prapengolahan Data	Penjelasan
1	Menghapus Nilai Kosong	Menghapus baris data yang memiliki nilai *null* pada salah satu atribut.
2	Cek dan Ubah Tipe Data	Memastikan semua kolom numerik bertipe *float* untuk proses komputasi.
3	Normalisasi Nilai Atribut	Menggunakan *Min-Max Scaling* agar seluruh atribut berada pada rentang 0–1.
4	Pemisahan Fitur dan Label	Memisahkan fitur (X) dan label target *potability* (y).
5	Pembagian *Data Train* dan *Tes*t	Data dibagi menjadi 80% data latih dan 20% data uji menggunakan *train_test_split*.

2.4. Pelatihan Model

Setelah proses prapengolahan data selesai, langkah selanjutnya adalah melakukan pelatihan model klasifikasi untuk menilai kelayakan air berdasarkan parameter-parameter kualitasnya. Dalam penelitian ini, algoritma *K-Nearest Neighbors* (KNN) dipilih karena kesederhanaannya dan efektivitasnya dalam mengklasifikasikan data berdasarkan kedekatan nilai fitur. Dataset yang telah dinormalisasi dibagi menjadi dua bagian: 80% sebagai data latih dan 20% sebagai data uji, menggunakan fungsi train_test_split() dari pustaka scikit-learn. Selanjutnya, dilakukan pengujian terhadap beberapa nilai K, yaitu K = 3, 5, dan 7.

2.5. Pengujian dan Evaluasi

Pengujian dan evaluasi dilakukan untuk menilai performa model *K-Nearest Neighbors* dalam mengklasifikasikan kelayakan air berdasarkan parameter kualitas seperti pH, TDS, dan *turbidity*. Dataset dibagi menjadi 80% untuk pelatihan dan 20% untuk pengujian menggunakan fungsi *train_test_split* dari pustaka scikit-learn, dengan tujuan memastikan kemampuan generalisasi model terhadap data baru. Evaluasi performa dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score*, yang memberikan gambaran menyeluruh mengenai efektivitas model dalam menangani data kelas seimbang. Selain itu, *confusion matrix* digunakan untuk mengilustrasikan jumlah prediksi benar dan salah pada masing-masing kelas (layak konsumsi dan tidak layak).

3. HASIL DAN PEMBAHASAN

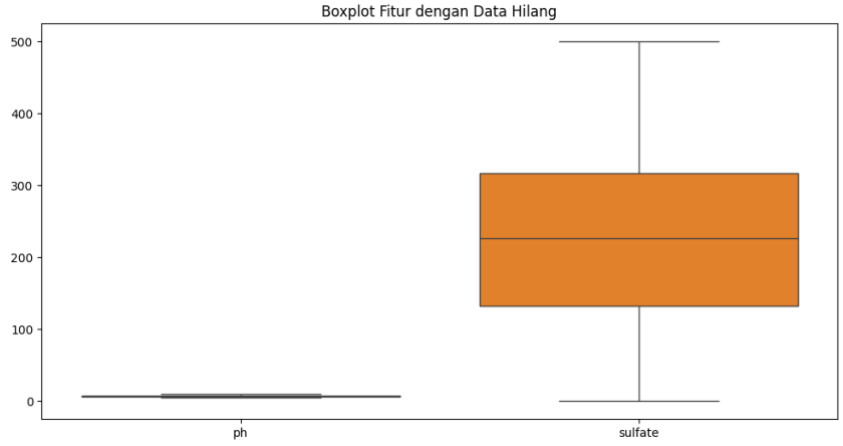
Bagian ini menyajikan hasil dari tahapan implementasi model serta analisis terhadap performa klasifikasi yang dilakukan. Proses dimulai dari prapengolahan data, pelatihan model menggunakan algoritma *K-Nearest Neighbors,* hingga evaluasi hasil prediksi menggunakan metrik klasifikasi. Setiap tahapan dijelaskan secara sistematis untuk menunjukkan bagaimana data diolah dan bagaimana model merespons pola yang terdapat dalam *dataset* kualitas air.

3.1. Prapengolahan Data

Prapengolahan data merupakan tahap krusial dalam proses analisis data karena kualitas data sangat menentukan akurasi dan keandalan model yang dibangun. Tahapan ini dilakukan dengan beberapa langkah sebagai berikut : Langkah pertama adalah mengimpor data menggunakan pustaka pandas dari berkas CSV yang berisi data kualitas air. Setelah data dimuat, dilakukan eksplorasi awal untuk memahami struktur data, jenis tipe data, dan mendeteksi adanya nilai kosong pada masing-masing atribut. Berdasarkan hasil inspeksi, diketahui bahwa atribut ph, tds, sulfate, dan conductivity mengandung nilai kosong (*missing values*). Untuk menangani hal ini, nilai kosong diisi menggunakan nilai median dari masing-masing kolom. Pemilihan median dilakukan karena sifatnya lebih tahan terhadap *outlier* dibandingkan dengan rata-rata.

Selanjutnya, dilakukan pengecekan dan penghapusan terhadap duplikat data untuk mencegah bias yang tidak diinginkan. Proses dilanjutkan dengan pembatasan terhadap nilai-nilai ekstrem yang tidak logis berdasarkan referensi standar kualitas air. Sebagai contoh, nilai ph dibatasi dalam rentang 0–14, *hardness* maksimal 500 mg/L CaCO3, tds maksimal 50.000 mg/L, chlorine maksimal 10 mg/L, dan batasan logis lainnya diterapkan untuk menjaga kualitas data.

Untuk mendeteksi *outlier* secara visual, digunakan *boxplot* pada atribut-atribut penting seperti ph dan sulfate. *Boxplot* ini membantu dalam memahami sebaran data dan mengidentifikasi nilai yang ekstrem atau mencurigakan.



Gambar 2 Boxplot

Distribusi kelas dari label target potability kemudian dianalisis. Terlihat bahwa data tidak seimbang, dengan label 0 (tidak layak minum) sebanyak 47.897 dan label 1 (layak minum) sebanyak 7.616. Ketidakseimbangan ini dapat memengaruhi performa model. Oleh karena itu, dilakukan teknik *undersampling* pada kelas mayoritas (label 0), dengan cara mengambil sampel acak sebanyak jumlah data pada kelas minoritas (label 1), yaitu 7.616 data. Kedua *subset* data kemudian digabung kembali, sehingga total data menjadi seimbang, masing-masing 7.616 untuk label 0 dan 1.

Setelah proses *undersampling*, dilakukan *shuffle* agar distribusi kelas tercampur secara acak dan tidak berpola. Label potability kemudian dipisahkan dari fitur untuk keperluan pelatihan model. Variabel target (y) adalah potability, sedangkan fitur (X) terdiri dari atribut-atribut yang relevan yaitu ph, sulfate, trihalomethanes, dan *turbidity*. Atribut lain seperti *temperature*, *hardness*, dan organic_carbon dihapus karena kontribusinya terhadap klasifikasi kualitas air tidak signifikan berdasarkan analisis awal.

Sebelum proses pelatihan dimulai, data dibagi menjadi dua *subset* yaitu data latih dan data uji dengan proporsi 80:20 menggunakan metode stratifikasi. Hasilnya, setiap *subset* memiliki distribusi kelas yang seimbang: data latih berisi 6.093 data untuk masing-masing kelas 0 dan 1, dan data uji masing-masing 1.523 data untuk kelas 0 dan 1.

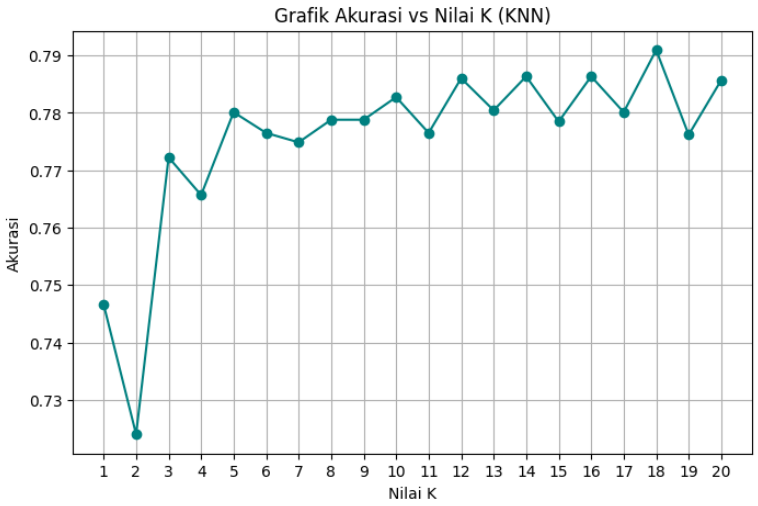
Tahap terakhir dari prapengolahan adalah normalisasi data menggunakan *Min-Max Scaling.* Proses ini mengubah nilai dari setiap fitur ke rentang 0 hingga 1 agar memiliki skala yang sama dan tidak mendominasi perhitungan jarak dalam algoritma *K-Nearest Neighbors* (KNN).

3.2. Pelatihan Model

Setelah proses normalisasi dan pembagian data menjadi data latih dan data uji, langkah berikutnya adalah pelatihan model menggunakan algoritma *K-Nearest Neighbors* (KNN). Algoritma KNN bekerja dengan cara mengklasifikasikan data baru berdasarkan mayoritas kelas dari k tetangga terdekatnya dalam ruang fitur. KNN dipilih karena kesederhanaannya serta kemampuannya yang baik dalam menyelesaikan masalah klasifikasi, terutama pada data yang memiliki dimensi tidak terlalu tinggi seperti pada kasus ini.

Pelatihan model dilakukan pertama kali dengan menggunakan nilai k sebesar 5. Setelah model dilatih pada data latih, data uji digunakan untuk mengevaluasi performa prediksi model. Evaluasi dilakukan dengan menggunakan metrik akurasi, serta classification report yang mencakup precision, recall, dan f1-score untuk masing-masing kelas, yaitu kelas 0 (tidak layak minum) dan kelas 1 (layak minum). Selain evaluasi awal tersebut, dilakukan pula eksperimen lanjutan untuk mengetahui pengaruh variasi nilai k terhadap performa model.

Eksperimen dilakukan dengan mencoba berbagai nilai k, mulai dari k = 1 hingga k = 20. Untuk setiap nilai k, model dilatih ulang dan diuji pada data uji yang sama, lalu akurasinya dicatat. Hasil dari eksperimen ini divisualisasikan dalam grafik berikut:



Gambar 3 Grafik Akurasi vs Nilai K pada KNN

Berdasarkan grafik tersebut, terlihat bahwa akurasi model sangat dipengaruhi oleh pemilihan nilai k. Nilai k yang sangat kecil, seperti k = 1 atau k = 2, cenderung menghasilkan akurasi yang lebih rendah, yaitu sekitar 72,5% pada k = 2. Hal ini kemungkinan disebabkan karena sensitivitas model terhadap *outlier* dan *noise* saat k terlalu kecil.

Sebaliknya, nilai k yang lebih besar cenderung memberikan hasil akurasi yang lebih stabil. Grafik menunjukkan bahwa mulai dari k = 5 ke atas, akurasi meningkat dan cenderung stabil meskipun terdapat sedikit fluktuasi. Nilai akurasi tertinggi tercapai pada k = 18, yaitu sebesar 79,1%, menjadikannya sebagai nilai k optimal untuk model ini. Dengan demikian, dapat disimpulkan bahwa pemilihan nilai k merupakan faktor krusial dalam algoritma KNN dan pada penelitian ini, k = 18 memberikan performa terbaik dalam memprediksi kelayakan air minum berdasarkan fitur-fitur seperti pH, kekeruhan, kadar sulfate, dan trihalomethanes.

3.3. Evaluasi Model

Pada tahap ini dilakukan evaluasi performa model klasifikasi *K-Nearest Neighbors* (KNN) terhadap data kualitas air yang telah melalui tahap praproses dan pelatihan model. Evaluasi dilakukan menggunakan beberapa metrik umum klasifikasi, yaitu akurasi, *precision, recall*, dan *f1-score*. Metrik ini dihitung berdasarkan hasil prediksi model terhadap data uji, dan disajikan dalam bentuk laporan klasifikasi serta *confusion matrix*.

Tabel 2 Confusion Matrix

	Prediksi Positif (1)	**Prediksi Negatif (0)**
Kenyataan Positif (1)	True Positive (TP)	False Negative (FN)
Kenyataan Negatif (0)	False Positive (FP)	True Negative (TN)

Evaluasi model klasifikasi umumnya didasarkan pada *confusion matrix* yang ada pada tabel 2, tabel ini menggambarkan jumlah prediksi yang benar dan salah untuk setiap kelas. Empat komponen penting dari *confusion matrix* adalah:

- **True Positive (TP)**: Prediksi positif dan kenyataannya positif.
- **True Negative (TN)**: Prediksi negatif dan kenyataannya negatif.
- **False Positive (FP)**: Prediksi positif tapi kenyataannya negatif.
- **False Negative (FN)**: Prediksi negatif tapi kenyataannya positif.

Gambar 4 menunjukkan hasil laporan klasifikasi yang diperoleh dari pengujian model KNN terhadap data uji sebanyak 3.047 data, yang terdiri dari dua kelas yaitu kelas 0 (air tidak layak minum) sebanyak 1.524 data dan kelas 1 (air layak minum) sebanyak 1.523 data. Model menghasilkan nilai akurasi sebesar 0.78, yang berarti bahwa 78% prediksi yang dilakukan oleh model sesuai dengan label yang sebenarnya. Pada kelas 0, *precision* mencapai 0.88 dan *recall* sebesar 0.65, menghasilkan f1-score sebesar 0.75. Sementara itu, pada kelas 1, *precision* sebesar 0.72 dan *recall* sebesar 0.91 menghasilkan f1-score sebesar 0.81. Hasil ini menunjukkan bahwa model lebih baik dalam mengenali air yang layak minum (kelas 1), ditunjukkan oleh nilai *recall* yang tinggi, meskipun ketepatan *(precision)* dalam memprediksi kelas tersebut masih perlu ditingkatkan.

```
Akurasi: 0.78

Laporan Klasifikasi:
              precision    recall  f1-score   support

           0       0.88      0.65      0.75      1524
           1       0.72      0.91      0.81      1523

    accuracy                           0.78      3047
   macro avg       0.80      0.78      0.78      3047
weighted avg       0.80      0.78      0.78      3047
```

Gambar 4 Hasil Evaluasi Model

Berdasarkan nilai-nilai tersebut, dilakukan perhitungan untuk memperoleh nilai TP, TN, FP, dan FN. Pada kelas 1 (air layak minum), *recall* sebesar 0.91 menunjukkan bahwa model berhasil mengenali sekitar 91% dari seluruh data yang memang layak, atau sekitar 1.385 data (TP). Dengan jumlah *support* sebesar 1.523 data untuk kelas 1, maka FN dapat dihitung sebesar $1.523 - 1.385 = 138$ data. Selanjutnya, *precision* sebesar 0.72 menunjukkan bahwa dari semua prediksi yang dinyatakan sebagai air layak, sekitar 72% benar. Dari nilai tersebut, FP diperkirakan sebesar 537 data. Akhirnya, dengan total data uji sebanyak 3.047, nilai TN dapat dihitung sebagai $3.047 - TP - FN - FP = 3.047 - 1.385 - 138 - 537 = 987$.

Precision merupakan ukuran yang menunjukkan seberapa akurat model dalam memprediksi kelas positif. Nilai ini dihitung dari proporsi prediksi positif yang benar (TP dibagi jumlah TP dan FP). *Recall*, di sisi lain, mengukur sensitivitas model terhadap kelas positif, yaitu seberapa besar model mampu mengenali data yang benar-benar positif (TP dibagi jumlah TP dan FN). F1-score merupakan rata-rata harmonik dari *precision* dan *recall*, yang memberikan keseimbangan antara keduanya. Nilai f1-score yang tinggi menandakan bahwa model tidak hanya akurat dalam prediksi positif, tetapi juga konsisten dalam mendeteksi seluruh data positif yang ada. Pada hasil evaluasi ini, f1-score makro dan rata-rata berbobot masing-masing sebesar 0.78, menunjukkan kinerja yang cukup stabil antara kedua kelas.

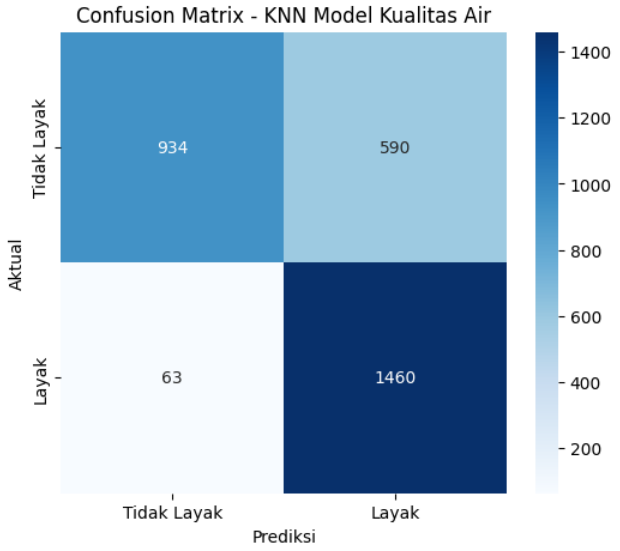
3.4. Pembahasan Hasil Evaluasi Model

Berdasarkan hasil evaluasi menggunakan *confusion matrix*, model *K-Nearest Neighbors* (KNN) yang dibangun menunjukkan performa yang baik dalam mengklasifikasikan kelayakan air minum. Seperti yang ditampilkan pada Gambar 5, model mampu memprediksi 934 data sebagai tidak layak secara benar (*True Negative*) dan 1.460 data sebagai layak secara benar (*True Positive*). Namun, terdapat 590 data tidak layak yang salah diprediksi sebagai layak (*False Positive*), serta 63 data layak yang salah diprediksi sebaagai tidak layak (*False Negative*).

Model menghasilkan akurasi sebesar 78%, yang mengindikasikan bahwa sekitar 78% dari keseluruhan prediksi yang dilakukan sesuai dengan label aktual. Nilai *precision* untuk kelas layak sebesar 72% menunjukkan bahwa dari seluruh data yang diprediksi sebagai air layak minum, sekitar 72% di antaranya benar-benar layak. Sementara itu, *recall* sebesar 91% mengindikasikan bahwa model mampu mendeteksi sebagian besar data yang memang layak minum, yakni sekitar 91%. Nilai f1-score untuk kelas layak sebesar 81% memperlihatkan adanya keseimbangan yang cukup baik antara ketepatan model dalam memprediksi dan kemampuannya dalam menangkap seluruh data positif.

Performa pada kelas tidak layak juga patut diperhatikan. Meskipun *precision*-nya tinggi, yaitu sebesar 88%, *recall*-nya lebih rendah, yakni 65%. Hal ini menunjukkan bahwa model lebih sering keliru dalam mengklasifikasikan air yang tidak layak sebagai layak, yang terlihat dari cukup besarnya jumlah *false positive*. F1-score pada kelas ini sebesar 75%, yang menunjukkan bahwa performa model masih cukup baik dalam mengenali air yang tidak layak, meskipun masih ada ruang untuk perbaikan.
Secara keseluruhan, hasil ini mencerminkan bahwa algoritma KNN mampu bekerja efektif dalam mengenali pola kualitas air berdasarkan fitur-fitur penting seperti pH, kekeruhan, kandungan sulfat, dan trihalomethanes. Dengan data yang telah melalui tahapan prapengolahan termasuk normalisasi dan pemilihan fitur, model menunjukkan performa yang cukup seimbang dalam

menangani kedua kelas. *Confusion matrix* yang dihasilkan menunjukkan bahwa mayoritas prediksi telah dilakukan secara tepat oleh model, menjadikan pendekatan ini menjanjikan untuk diterapkan dalam sistem pemantauan kualitas air secara otomatis dan efisien. *Confusion matrix* yang dihasilkan pada Gambar, menunjukkan bahwa prediksi untuk kedua kelas, baik layak maupun tidak layak, telah dilakukan secara tepat oleh model dalam sebagian besar kasus. Kemampuan ini memberikan potensi besar bagi penerapan model KNN dalam mendukung sistem pemantauan kualitas air secara otomatis dan cepat.



Gambar 5 Hasil Confusion Matrix

4. KESIMPULAN

Berdasarkan dari hasil penelitian ini menunjukkan bahwa model klasifikasi kualitas air menggunakan algoritma *K-Nearest Neighbors* (KNN) yang dikembangkan berhasil mencapai hasil yang andal dan akurat, sesuai dengan tujuan awal penelitian. Proses prapengolahan data, pemilihan fitur penting seperti pH, kekeruhan, sulfate, dan trihalomethanes, serta pelatihan dan evaluasi model, membuktikan bahwa KNN mampu memberikan performa yang baik dengan akurasi mencapai 78%. Evaluasi menggunakan *confusion matrix* juga mengonfirmasi kemampuan model dalam mengenali pola kualitas air secara konsisten, terutama pada kelas air layak minum dengan keseimbangan *precision* dan *recall* yang memadai.

Model ini berpotensi untuk diterapkan dalam sistem *monitoring* kualitas air secara *real-time,* khususnya jika dikombinasikan dengan teknologi sensor dan *Internet of Things* (IoT). Pengembangan di masa depan dapat difokuskan pada penerapan algoritma yang lebih kompleks seperti Random Forest atau Neural Network, serta peningkatan kualitas data dengan menambahkan parameter kimiawi dan biologis yang relevan. Dengan demikian, sistem klasifikasi ini tidak hanya berfungsi sebagai alat bantu analisis, tetapi juga dapat menjadi sistem peringatan dini untuk menjaga kesehatan dan keselamatan masyarakat.

REFERENSI

[1] Nurmahaludin, "Klasifikasi Kualitas Air Pdam Menggunakan Algoritma Knn Dan K-means," *Klasifikasi Kualitas Air Pdam Menggunakan Algoritma Knn Dan K-means*, vol. 1, no. 1, 2019.
[2] A. Muhtar , P. Wibawa , and M. Kallista, "Klasifikasi Kualitas Sungai Air Menggunakan Metode Pembelajaran Mesin k-Nearest Neighbour," *Klasifikasi Kualitas Sungai Air Menggunakan Metode Pembelajaran Mesin k-Nearest Neighbour*, vol. 11, no. 1, Feb. 2024.
[3] M. Syarifuddin, "Klasifikasi Kualitas Air Pada Program Penyediaan Air Minum Dan Sanitasi Berbasis Masyarakat Desa Semenpinggir Dengan Metode Algoritma K-Nearest Neighbor," *Klasifikasi Kualitas Air Pada Program Penyediaan Air Minum Dan Sanitasi Berbasis Masyarakat Desa Semenpinggir Dengan Metode Algoritma K-Nearest Neighbor*, vol. 2, no. 1, Mar. 2024.
[4] United States, "The Sustainable Development Goals Report," *The Sustainable Development Goals Report*, vol. 1, no. 1, 2021.
[5] World Health Organization, "Progress on household drinking water, sanitation and hygiene, 2000-2020: Five years into the SDGs," Unicef Data. [Online]. Available: https://data.unicef.org/resources/progress-on-household-drinking-water-sanitation-and-hygiene-2000-2020/
[6] World Health Organization: WHO, "Drinking-water," *World Health Organization: WHO*, Sep. 13, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/drinking-water
[7] P. Sawant, "Physico-chemical parameters for testing of water – A review," *Physico-chemical parameters for testing of water – A review*, vol. 3, no. 3, 2012.
[8] S. Putri, "Penerapan Metode SVM pada Klasifikasi Kualitas Air," *Penerapan Metode SVM pada Klasifikasi Kualitas Air*, vol. 3, no. 2, 2023.
[9] Q. Jemila, Dhanalakshmi, and Amutha, "Water Quality Prediction Using Decision Tree and KNN," *Water Quality Prediction Using Decision Tree and KNN*, vol. 9, no. 1, Jan. 2024.
[10] M. Hasin, "Penerapan Neural Network sebagai Klasifikasi Kualitas Air Hasil Filtrasi Reverse Osmosis," *Penerapan Neural Network sebagai Klasifikasi Kualitas Air Hasil Filtrasi Reverse Osmosis*, vol. 11, no. 3, Sep. 2024.

[11] Kemenkes, *Profil Kesehatan Indonesia*. Kementrian Kesehatan, 2022.
[12] T. Brian, "Application of K-Nearest Neighbor (KNN) Algorithm to Predict Drinking Water Quality," *Application of K-Nearest Neighbor (KNN) Algorithm to Predict Drinking Water Quality*, vol. 5, no. 1, Jan. 2025.
[13] S. Ulum, R. F. Alifa, P. Rizkika, and C. Rozikin, "Perbandingan Performa Algoritma KNN dan SVM dalam Klasifikasi Kelayakan Air Minum," *Generation Journal*, vol. 7, no. 2, Jul. 2023.
[14] Pangaribuan, J. J., Maulana, A., & Romindo, R. (2024). UNLEASHING THE POWER OF SVM AND KNN: ENHANCED EARLY DETECTION OF HEART DISEASE. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer), 10(2), 342-351.
[15] S. Jadhav, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *International Journal of Science and Research (IJSR)*, vol. 5, no. 1, 2016.
[16] A. Kumar, "Review on Data Mining Techniques for Prediction of Water Quality," *Review on Data Mining Techniques for Prediction of Water Quality*, vol. 6, no. 6, Jun. 2019.
[17] P. Padmaja, "Water Quality Prediction Using Machine Learning Algorithms," *Water Quality Prediction Using Machine Learning Algorithms*, vol. 10, no. 4, Apr. 2023.