Contents lists available at MID Publisher International

Technology Sciences Insights Journal

Journal homepage: https://journal.midpublisher.com/index.php/tsij

Air Quality Monitoring System Metode KNN (K-Nearest Neighbor) untuk Mendeteksi Kondisi Kualitas Air

Ardiyan Syah Ariansyah
Politeknik Negeri Batam, Batam, Indonesia

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima: Juni 2025
Revisi: Juni 2025
Diterima: Juli 2025
Dipublikasi: Juli 2025

Kata Kunci:
Water Quality; K-Nearest Neighbor; Machine Learning; Classification; Monitoring System

*Penulis Korespondensi:
syahardi132@gmail.com

ABSTRAK

Penelitian ini bertujuan untuk membuat sebuah system yang berguna untuk mempermudah manusia untuk mengetahui kondisi kualitas air secara manual menggunakan metode KNN (K-Nearest Neighbor). Kualitas air adalah aspek penting dalam kehidupan sehari-hari, dan upaya untuk memastikan air yang aman untuk dikonsumsi merupakan prioritas. Penelitian ini mengajukan pertanyaan utama: "Apakah metode KNN dapat memberikan pendekatan yang efektif dalam mengklasifikasikan kualitas air?" Metodologi penelitian ini melibatkan penggunaan dataset sekunder mengenai kualitas air dan menerapkan algoritma KNN untuk mengkategorikan data tersebut. Hasil dari metode KNN dibandingkan dengan metode klasifikasi lain yang umum digunakan dalam analisis kualitas air. Penelitian ini bertujuan untuk memberikan pandangan yang lebih dalam dan pemahaman yang lebih baik tentang potensi metode KNN. Dengan memakai metode KNN (KNearest Neighbor) diperoleh nilai ketelitian dari K, nilai K yang menggunakan K=5 akurasinya sebesar 97%, setelah KNN mendapatkan hasil dari deteksi, hasil tersebut akan berguna untuk mengatur tindakan dari aktuator. Hasil penelitian ini diharapkan dapat memberikan wawasan baru tentang efektivitas metode KNN dalam mengklasifikasikan kualitas air yang dapat membantu masyarakat dan pemerintah dalam mengambil tindakan respons yang lebih efisien terkait dengan potensi kontaminasi dan masalah kualitas air lainnya. Dengan demikian, penelitian ini memiliki potensi untuk memberikan kontribusi berharga pada pemahaman kita tentang kualitas air dan pemanfaatan metode KNN dalam analisisnya.

ABSTRACT

This research aims to create a system that is useful to make it easier for humans to know the condition of water quality manually using the KNN (K-Nearest Neighbor) method. Water quality is an important aspect of daily life, and efforts to ensure safe water for consumption are a priority. This research asks the main question: "Can the KNN method provide an effective approach in classifying water quality?" The research methodology involved using a secondary dataset on water quality and applying the KNN algorithm to categorize the data. The results of the KNN method were compared with other classification methods commonly used in water quality analysis. This research aims to provide a deeper look and better understanding of the potential of the KNN method. By using the KNN (K-Nearest Neighbor) method, the accuracy value of K is obtained, the value of K using K = 5 is 97% accuracy, after KNN gets the results of detection, these results will be useful for regulating the actions of the actuator. The results of this study are expected to provide new insights into the effectiveness of the KNN method in classifying water quality that can assist the public and government in taking more efficient response actions related to potential contamination and other

water quality issues. As such, this research has the potential to make valuable contributions to our understanding of water quality and the utilization of the KNN method in its analysis.

PENDAHULUAN

Air adalah unsur yang mendominasi wilayah terluas di planet bumi, mencakup hampir 71% dari seluruh permukaan dan ada di berbagai bentuk, termasuk air laut, air sungai, air permukaan, air atmosfer, air rawa, dan air tanah. Selain memenuhi kebutuhan dasar manusia seperti konsumsi dan memasak, air juga memiliki peran penting dalam industri, pertanian, dan pembangunan infrastruktur. Kualitas air memiliki dampak yang sangat signifikan pada kehidupan kita. Faktor-faktor lingkungan fisik, seperti pola curah hujan, topografi, dan jenis batuan, berperan dalam menentukan kuantitas air di suatu wilayah. Sementara itu, kualitas air sangat dipengaruhi oleh aspek-aspek sosial seperti kepadatan penduduk dan tingkat interaksi sosial. Di tengah kompleksitas ini, perhatian terhadap kualitas air menjadi semakin penting. Pembangunan yang pesat seringkali mengubah aliran sungai menjadi area permukiman dan industri di kota-kota, yang pada gilirannya dapat mengancam kualitas air yang mengalir di sepanjang sungai tersebut.

Penelitian tentang klasifikasi air minum telah banyak dilakukan, terutama mengenai klasifikasi kualitas air minum menggunakan machine learning. Penelitian oleh Aldi dkk. (2023) menggunakan metode Naive Bayes, Decision Tree, dan K-Nearest Neighbors untuk menemukan metode yang paling akurat. Hasil penelitian menunjukkan bahwa metode Decision Tree memiliki akurasi tertinggi. Penelitian oleh Prismahardi dkk. (2023) menggunakan metode Support Vector Machine, Decision Tree, Naïve Bayes, dan Artificial Neural Network. Hasil penelitian menunjukkan bahwa metode Random Forest Classifier memiliki akurasi tertinggi. Kemudian penelitian yang dilakukan oleh Ilyas dkk yang bertujuan untuk mengklasifikasikan kualitas air minum menggunakan algoritma Random Forest. Data yang digunakan adalah data kualitas air minum dari 267 sampel air sumur di Jakarta. Hasil penelitian menunjukkan bahwa algoritma Random Forest dapat mengklasifikasikan kualitas air minum dengan akurasi sebesar 82,3%.

Pada penelitian kali ini bertujuan untuk memberikan hal inovasi dengan menggunakan metode K-Nearest Neighbor (KNN) dalam menentukan kualitas air. Efektivitas metode KNN akan dievaluasi dalam mengklasifikasikan data kualitas air yang diharapkan dapat memberikan wawasan yang lebih baik dan membantu dalam mengatasi tantangan kualitas air yang dihadapi oleh masyarakat Indonesia dan negara-negara lainnya. KNN adalah algoritma pembelajaran mesin terawasi yang digunakan untuk klasifikasi dan regresi, memanipulasi data training serta mengklasifikasikan data testing berdasarkan metrik jarak. Dengan mencari K tetangga terdekat dari data uji, algoritme ini melakukan klasifikasi berdasarkan mayoritas label kelas, dan termasuk dalam kelas instance base learning serta lazy learning. Dengan penerapan teknik K-Nearest Neighbor (KNN), merupakan salah satu pendekatan klasifikasi terhadap kumpulan data yang berfokus pada mayoritas kategori. Tujuannya adalah untuk mengategorikan objek baru dengan merujuk pada atribut dan contohcontoh sampel dari data pelatihan. Dengan demikian, upaya ini bertujuan untuk mendekati akurasi yang lebih tinggi saat melakukan evaluasi pembelajaran.

LANDASAN TEORI (Optional)

1. KNN (K-Nearest Neighbor)

K-Nearst Neighbor (KNN) adalah suatu algoritma yang digunakan untuk mengklasifikasikan data dengan memanfaatkan data latih (training record) dari tetangga terdekat, di mana k merupakan jumlah tetangga terdekat yang digunakan dalam proses klasifikasi. KNN adalah jenis algoritma yang sederhana dibandingkan algoritma lain dalam

machine learning. Prinsip kerjanya adalah dengan menghitung jarak perbedaan antara data uji dan data latih yang ada dalam sistem, kemudian mencari nilai terkecil atau paling mirip untuk mengklasifikasikan data tersebut. Suatu Objek di klasfikasikan dengan class mayoritas dari class tetangga, dimana diambil class yang paling banyak muncul dalam batasan K tertentu dari tetangganya,sehingga diperoleh class baru sesuai dengan tetangga yang paling umum, jika K=1, maka objek ditetapkan sesuai dengan class tetangga terdekatnya. Pada fase klasifikasi, K adalah sebuah konstanta yang ditetapkan pengguna. Class baru dari suatu data test diklasifikasikan dengan menetapkan class yang paling sering muncul diantara sampel ke titik data training yang ada. Metode K-Nearst Neighbor pada prinsip kerjanya mencari jarak terdekat antara data yang akan dievaluasi dengan K tetangga (Neighbor) terdekatnya dalam data sampel.

Euclidean distance merupakan jarak antara dua titik atau koordinat yang dihitung menggunakan rumus Pythagoras. Ini adalah panjang garis lurus yang menghubungkan dua titik dalam ruang, yang disebut titik a dan titik b. Garis ini, juga dikenal sebagai garis miring, terbentang di antara sumbu x dan sumbu y dengan koordinat yang diberikan untuk titik a dan titik b. Gambar berikut menunjukan rumus perhitungan untuk mencari jarak antara dua titik yaitu titik pada data training (x) dan titik pada data testing (y) maka digunakan rumus Euclidean, seperti pada persamaan berikut.

$$\mathrm{D(x,y)}\ \sqrt{\sum_{k=1}^{n}(p_1-q_1)^2}$$

Gambar 1. Rumus Euclidean

Dengan D adalah jarak antara titik pada data training x dan titik data testing yang akan diklasifikasi, dimana x=x1, x2, …, xi dan y1, y2, … yi dan i mempresentasikan nilai atribut serta n merupakan dimensi atribut.

a) Menentukan parameter K (jumlah tetangga paling dekat).
b) Menghitung kuadrat jarak Euclid (query instance) masing-masing objek terhadap data sampel yang diberikan.
c) Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak Euclid terkecil.
d) Mengumpulkan kategori Y (klasifikasi nearest neighbor).
e) Dengan menggunakan kategori Nearest Neighbor yang paling mayoritas maka dapat diprediksi nilai query instance yang telah dihitung.

2. Datashet

Dataset adalah kumpulan data yang terdiri dari beberapa entitas atau sampel yang diambil dari berbagai sumber. Dalam konteks penelitian ilmiah, dataset biasanya berisi sekumpulan variabel atau atribut yang diukur atau diamati pada setiap entitas, serta informasi tambahan yang relevan untuk analisis. Dataset merupakan komponen penting dalam penelitian karena menjadi dasar untuk melakukan analisis, pembuatan model, dan menyimpulkan temuan penelitian. Dalam penelitian mengenai metode KNN untuk menentukan kualitas air, dataset berisi data pengukuran atau observasi berbagai parameter kualitas air, seperti konsentrasi logam berat, kandungan bakteri, dan kualitas kimia lainnya. Setiap baris dalam dataset mewakili satu entitas atau contoh, seperti sampel air yang diambil dari berbagai lokasi atau sumber. Sedangkan, setiap kolom merepresentasikan atribut atau parameter yang diamati pada setiap entitas.

Contoh dataset yang digunakan dalam penelitian ini diambil dari website Kaggle.com yang berjudul waterQuality1. Data ini telah dikumpulkan dari berbagai lokasi dan telah melalui proses validasi dan pengolahan sebelum digunakan dalam klasifikasi. Sedangkan klasifikasi sendiri merupakan langkah dimana data dikelompokkan ke dalam kelas-kelas yang telah terdefinisi sebelumnya, berdasarkan atribut atau fitur-fitur tertentu yang dimiliki oleh data

tersebut. Proses ini bertujuan untuk mengidentifikasi pola atau karakteristik yang membedakan setiap kelas dan memungkinkan untuk pengambilan keputusan yang lebih baik berdasarkan pada klasifikasi tersebut

Klasifikasi adalah tahap fundamental dalam analisis data dan pembelajaran mesin, di mana model yang telah dihasilkan dari data latih digunakan untuk memprediksi kelas dari data uji yang belum pernah dilihat sebelumnya. Dalam konteks penelitian dan aplikasi praktis, klasifikasi dapat digunakan dalam berbagai bidang seperti pengenalan pola, diagnosis medis, analisis risiko keuangan, dan banyak lagi. Proses ini melibatkan penggunaan algoritma pembelajaran mesin, seperti K-Nearest Neighbor untuk memetakan data ke dalam kelaskelas yang telah ditentukan sebelumnya, sehingga memberikan pemahaman yang lebih dalam tentang karakteristik dan pola yang ada dalam data tersebut.

3. Kualitas Air

Kualitas air merupakan ukuran kondisi fisik, kimia, dan biologis air yang menentukan kelayakannya untuk digunakan dalam kehidupan sehari-hari, baik untuk konsumsi, pertanian, maupun keperluan industri. Air yang memiliki kualitas baik harus memenuhi standar tertentu agar tidak membahayakan kesehatan manusia dan lingkungan. Parameter kualitas air umumnya meliputi kandungan zat kimia, tingkat kekeruhan, kadar gas terlarut, serta keberadaan mikroorganisme berbahaya.

Penurunan kualitas air dapat disebabkan oleh berbagai faktor, seperti aktivitas industri, limbah domestik, pertanian, dan pertumbuhan populasi yang tidak terkendali. Oleh karena itu, pemantauan kualitas air secara berkelanjutan menjadi sangat penting untuk mendeteksi pencemaran sejak dini dan mencegah dampak negatif yang lebih luas. Dengan berkembangnya teknologi, pemantauan kualitas air kini dapat dilakukan secara otomatis dengan bantuan sistem komputasi dan metode kecerdasan buatan

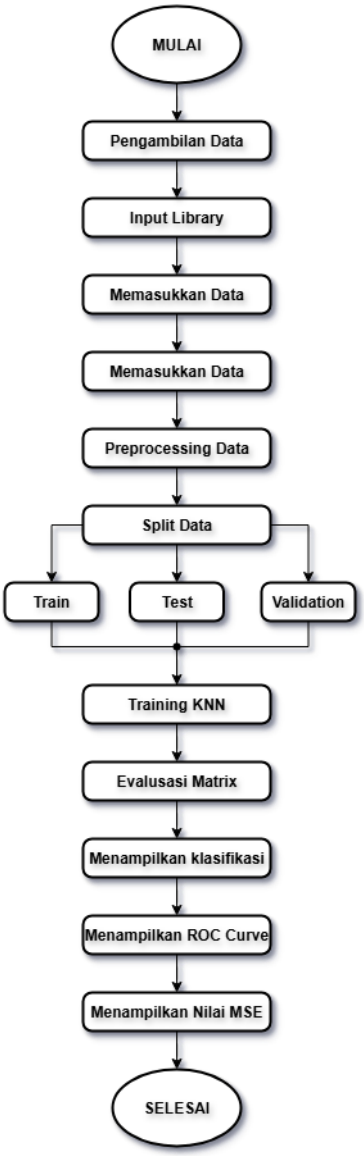
4. Machine Learning

Machine Learning merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem komputer untuk belajar dari data tanpa harus diprogram secara eksplisit. Machine learning bekerja dengan mengenali pola dari data latih (training data) dan kemudian menggunakan pola tersebut untuk melakukan prediksi atau klasifikasi terhadap data baru.

Dalam konteks analisis kualitas air, machine learning digunakan untuk mengolah data pengukuran berbagai parameter kualitas air dan mengklasifikasikannya ke dalam kategori tertentu, seperti layak atau tidak layak konsumsi. Metode ini dinilai efektif karena mampu menangani data dalam jumlah besar serta memberikan hasil analisis yang cepat dan akurat..

METODE PENELITIAN

Penelitian ini dimulai dengan mengumpulkan dataset yang diperoleh dari situs UC Irvine Machine Learning, yang merupakan salah satu basis data dalam kompetisi di bidang penglihatan komputer dan dapat diakses oleh publik dengan judul AirQualityUCI dalam bentuk extensi .csv. Data tersebut kemudian diproses secara preprocessing untuk memastikan bahwa data siap digunakan sebagai data target. Total data target yang terkumpul adalah 9358 record. Proses klasifikasi melibatkan pencarian model atau fungsi yang dapat menggambarkan serta membedakan antara konsep atau kelas-kelas data. Tujuannya adalah untuk memprediksi kelas dari objek yang tidak memiliki label, sehingga memberikan perkiraan mengenai kelasnya. Data latih selanjutnya dimasukkan ke dalam algoritma K-Nearest Neighbor untuk mengembangkan model klasifikasi. Setelah model klasifikasi terbentuk, dilakukan uji akurasi menggunakan data uji. Gambar 1 menunjukkan alur dari model klasifikasi yang dibangun menggunakan K-Nearest Neighbor. Metode ini digunakan untuk mencari tetangga terdekat dari data uji dan melakukan klasifikasi berdasarkan mayoritas label kelas dari tetangga terdekat tersebut.



Gambar 2. Tahapan mode KNN

PEMBAHASAN

Penelitian ini bertujuan untuk mengembangkan sebuah program yang dapat melakukan penentuan kualitas air dengan menggunakan Metode K-Nearest Neighbor (KNN). Program ini menggunakan dataset yang diperoleh dari UC Irvine Machine Learning, dengan judul " AirQualityUCI". Dataset ini mengandung informasi mengenai kualitas air yang baik untuk digunakan dalam sebuah pekerjaan industry dan datashet yang diambil dari beberapa jenis kualitas air digunakan sebagai fitur-fitur klasifikasi untuk menentukan apakah suatu sampel air layak konsumsi atau tidak. Setiap unsur memiliki nilai-nilai yang memiliki batasan yang menandakan apakah nilainya berada dalam kisaran yang aman untuk dikonsumsi. Dalam konteks ini, batasan-batasan tersebut dapat diartikan sebagai nilai-nilai ambang yang harus dipatuhi oleh setiap unsur agar air dianggap aman untuk dikonsumsi.

Berikut ini merupakan gambaran sebagian dari tampilan dataset yang diperoleh dari sumbernya, yaitu situs UC Irvine Machine Learning, yang berjudul " AirQualityUCI ". Dataset ini mengandung sebanyak 9358 sampel yang beragam, namun untuk tujuan penulisan artikel ini, tidak semua sampel akan ditampilkan secara lengkap. Meskipun demikian, bagian dari dataset yang disajikan akan memberikan pandangan awal mengenai isi dan komposisi data yang digunakan dalam penelitian ini.

Table 1. Dataset hasil uji

3	10/03/2004	19:00:00	2	1292	112	9,4	955	103	1174	92	1559	972	13,3	47,7	0,7255
4	10/03/2004	20:00:00	2,2	1402	88	9,0	939	131	1140	114	1555	1074	11,9	54,0	0,7502
5	10/03/2004	21:00:00	2,2	1376	80	9,2	948	172	1092	122	1584	1203	11,0	60,0	0,7867
6	10/03/2004	22:00:00	1,6	1272	51	6,5	836	131	1205	116	1490	1110	11,2	59,6	0,7888
7	10/03/2004	23:00:00	1,2	1197	38	4,7	750	89	1337	96	1393	949	11,2	59,2	0,7848
8	11/03/2004	00:00:00	1,2	1185	31	3,6	690	62	1462	77	1333	733	11,3	56,8	0,7603
9	11/03/2004	01:00:00	1	1136	31	3,3	672	62	1453	76	1333	730	10,7	60,0	0,7702
10	11/03/2004	02:00:00	0,9	1094	24	2,3	609	45	1579	60	1276	620	10,7	59,7	0,7648
11	11/03/2004	03:00:00	0,6	1010	19	1,7	561	-200	1705	-200	1235	501	10,3	60,2	0,7517
12	11/03/2004	04:00:00	-200	1011	14	1,3	527	21	1818	34	1197	445	10,1	60,5	0,7465
13	11/03/2004	05:00:00	0,7	1066	8	1,1	512	16	1918	28	1182	422	11,0	56,2	0,7366
14	11/03/2004	06:00:00	0,7	1052	16	1,6	553	34	1738	48	1221	472	10,5	58,1	0,7353
15	11/03/2004	07:00:00	1,1	1144	29	3,2	667	98	1490	82	1339	730	10,2	59,6	0,7417
16	11/03/2004	08:00:00	2	1333	64	8,0	900	174	1136	112	1517	1102	10,8	57,4	0,7408
17	11/03/2004	09:00:00	2,2	1351	87	9,5	960	129	1079	101	1583	1028	10,5	60,6	0,7691
18	11/03/2004	10:00:00	1,7	1233	77	6,3	827	112	1218	98	1446	860	10,8	58,4	0,7552
19	11/03/2004	11:00:00	1,5	1179	43	5,0	762	95	1328	92	1362	671	10,5	57,9	0,7352
20	11/03/2004	12:00:00	1,6	1236	61	5,2	774	104	1301	95	1401	664	9,5	66,8	0,7951
21	11/03/2004	13:00:00	1,9	1286	63	7,3	869	146	1162	112	1537	799	8,3	76,4	0,8393
22	11/03/2004	14:00:00	2,9	1371	164	11,5	1034	207	983	128	1730	1037	8,0	81,1	0,8736
23	11/03/2004	15:00:00	2,2	1310	79	8,8	933	184	1082	126	1647	946	8,3	79,8	0,8778
24	11/03/2004	16:00:00	2,2	1292	95	8,3	912	193	1103	131	1591	957	9,7	71,2	0,8569
25	11/03/2004	17:00:00	2,9	1383	150	11,2	1020	243	1008	135	1719	1104	9,8	67,6	0,8185
26	11/03/2004	18:00:00	4,8	1581	307	20,8	1319	281	799	151	2083	1409	10,3	64,2	0,8065
27	11/03/2004	19:00:00	6,9	1776	461	27,4	1488	383	702	172	2333	1704	9,7	69,3	0,8319
28	11/03/2004	20:00:00	6,1	1640	401	24,0	1404	351	743	165	2191	1654	9,6	67,8	0,8133
29	11/03/2004	21:00:00	3,9	1313	197	12,8	1076	240	957	136	1707	1285	9,1	64,0	0,7419
30	11/03/2004	22:00:00	1,5	965	61	4,7	749	94	1325	85	1333	821	8,2	63,4	0,6905
31	11/03/2004	23:00:00	1	913	26	2,6	629	47	1565	53	1252	552	8,2	60,8	0,6657

Data tersebut dikelompokkan ke dalam kategori klasifikasi 1, yang mengindikasikan bahwa data tersebut dianggap aman atau memenuhi standar yang diperlukan untuk dapat dikonsumsi.

HASIL

Berikut implementasi klasifikasi K-Nearest Neighbors (KNN) menggunakan Gogole Colab yang melibatkan beberapa langkah penting, seperti yang dijelaskan berikut:

a) Import Library

```
# === 1. Import Library ===
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix, roc_curve, auc, mean_squared_error
```

Gambar 3. Proses Import Library

Pada cell pertama, program mengimpor berbagai pustaka yang dibutuhkan untuk proses analisis dan pemodelan data. pandas dan numpy digunakan untuk membaca dan memproses data dalam bentuk tabel dan numerik. Kemudian, matplotlib.pyplot dan seaborn digunakan untuk visualisasi data dan hasil evaluasi model. Dari library scikit-learn, diimpor berbagai modul penting seperti train_test_split untuk membagi data, StandardScaler untuk normalisasi, KNeighborsClassifier sebagai algoritma utama, serta berbagai metrik evaluasi seperti classification_report, confusion_matrix, roc_curve, auc, dan mean_squared_error.

b) Load Data & Buat Label Klasifikasi

```
# === 2. Load Data ===
df = pd.read_csv("AirQualityUCI.csv", sep=';', decimal=',')
df = df.dropna(how='all', axis=1)  # hapus kolom kosong
df = df.dropna(how='any')          # hapus baris yang ada NaN
df = df.select_dtypes(include='number')  # ambil hanya data numerik

# Hapus nilai error dari CO(GT)
df = df[df['CO(GT)'] != -200]

# === 3. Buat Label Klasifikasi ===
df['Label'] = df['CO(GT)'].apply(lambda x: 1 if x > 2 else 0)  # threshold 2 mg/m³
```

Gambar 4. Proses Load Data dan Melakukan Labelling Data

Pada bagian ini, dataset dibaca dari file AirQualityUCI.csv menggunakan pandas dengan penyesuaian terhadap format file (separator; dan desimal ,). Setelah itu, program membersihkan data dengan menghapus kolom kosong dan baris yang mengandung nilai NaN. Selanjutnya, hanya data numerik yang diambil untuk proses analisis. Salah satu kolom penting, CO(GT), dibersihkan dari nilai error yaitu -200. Terakhir, dibuatlah kolom Label untuk klasifikasi kualitas udara, di mana kadar CO di atas 2 mg/m³ diberi label 1 (tidak sehat) dan sisanya diberi label 0 (sehat)

c) Fitur & Label

```
# === 4. Fitur & Label ===
X = df.drop(columns=['Label'])
y = df['Label']

# Scaling karena KNN sensitif terhadap jarak
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Split data
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42, stratify=y
)
```

Gambar 5. Proses Scalling Data

Cell ketiga memisahkan antara fitur (X) dan target klasifikasi (y atau Label). Karena algoritma KNN sangat bergantung pada perhitungan jarak antar titik data, dilakukan normalisasi atau scaling terhadap fitur menggunakan StandardScaler agar semua fitur berada pada skala yang sama. Setelah normalisasi, data dibagi menjadi dua bagian: data pelatihan sebanyak 80% dan data pengujian sebanyak 20%. Pembagian ini dilakukan dengan stratifikasi agar distribusi label tetap seimbang di kedua subset data.

d) Tarining KNN

```
# === 5. Training KNN ===
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)

# Prediksi
y_pred = knn.predict(X_test)
```

Gambar 6. Proses training menggunakan KNN

Cell ini merupakan tahap pelatihan model menggunakan algoritma K-Nearest Neighbors (KNN). Model KNN diinisialisasi dengan jumlah tetangga k = 5, lalu dilatih (fit) menggunakan data pelatihan yang telah di-normalisasi. Setelah proses training selesai, model digunakan untuk memprediksi hasil klasifikasi terhadap data uji (X_test). Hasil prediksi disimpan dalam variabel y_pred.

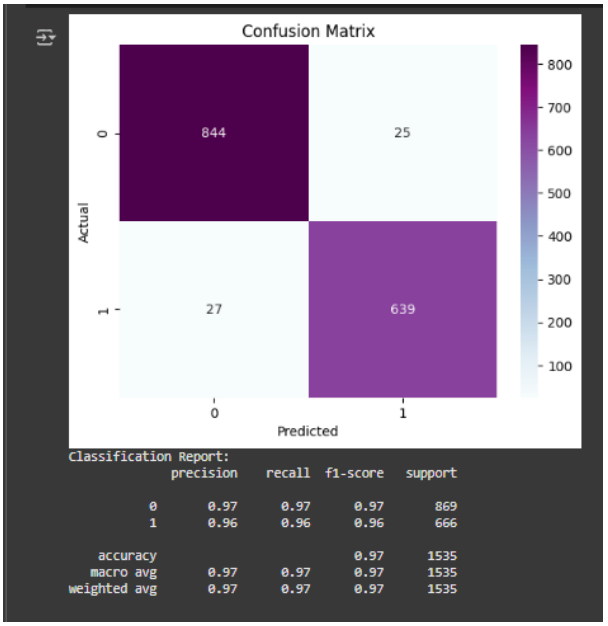
e) Confiusion Matrix

```
# === 6. Confusion Matrix ===
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, cmap='BuPu', fmt='d')
plt.title("Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.show()

# Classification Report
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

Gambar 7. Proses pemanggilan confusion matrix

Pada cell ini dilakukan evaluasi kinerja model KNN menggunakan confusion matrix dan classification report. Confusion matrix dihitung untuk membandingkan hasil prediksi dengan label aktual (y_test) dan divisualisasikan dalam bentuk heatmap agar lebih mudah dipahami. Setelah itu, dicetak classification report yang memuat metrik evaluasi seperti precision, recall, f1-score, dan akurasi untuk masing-masing kelas (layak dan tidak layak). Evaluasi ini bertujuan untuk melihat sejauh mana model dapat mengklasifikasikan kualitas udara dengan benar.



Classification Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	869
1	0.96	0.96	0.96	666
accuracy			0.97	1535
macro avg	0.97	0.97	0.97	1535
weighted avg	0.97	0.97	0.97	1535

Gambar 8. Hasil confusion matrix dan matrix evaluasi

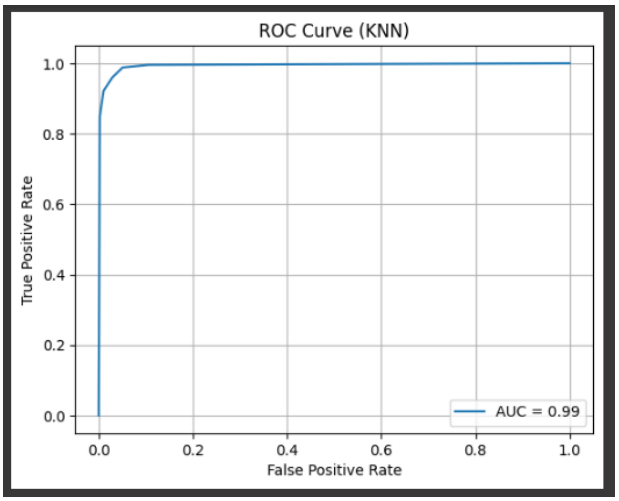
f) ROC Curve

```
# === 7. ROC Curve ===
y_prob = knn.predict_proba(X_test)[:, 1]
fpr, tpr, _ = roc_curve(y_test, y_prob)
roc_auc = auc(fpr, tpr)

plt.plot(fpr, tpr, label=f"AUC = {roc_auc:.2f}")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve (KNN)")
plt.legend()
plt.grid()
plt.show()
```

Gambar 9. Mencari ROC Curve

Di bagian ini, program menghitung dan memvisualisasikan kurva ROC (Receiver Operating Characteristic), yang menggambarkan kemampuan model dalam membedakan antara dua kelas. Probabilitas hasil prediksi dari model KNN diambil menggunakan predict_proba, lalu dihitung nilai false positive rate (FPR) dan true positive rate (TPR). Dari grafik ROC ini juga dihitung nilai AUC (Area Under Curve) yang menunjukkan seberapa baik model dalam klasifikasi secara umum — semakin mendekati 1, semakin baik.



Gambar 10. Hasil ROC Curve

g) Mean Squared Error (MSE)

```
# === 8. Mean Squared Error ===
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error (MSE): {mse:.4f}")

# Cross Validation
cv_scores = cross_val_score(knn, X_scaled, y, cv=5)
print("Cross-validation scores:", cv_scores)
print("Average CV score:", np.mean(cv_scores))
```

Gambar 11. Proses mencari MSE

Pada bagian ini, evaluasi model KNN dilanjutkan dengan menghitung Mean Squared Error (MSE), yaitu rata-rata kuadrat dari selisih antara label aktual dan hasil prediksi. Meskipun MSE lebih umum digunakan pada regresi, di sini digunakan sebagai tambahan metrik numerik untuk melihat seberapa jauh kesalahan model dalam klasifikasi biner. Setelah itu, dilakukan cross-validation sebanyak 5 kali lipat (5-fold cross-validation) terhadap seluruh data terstandardisasi untuk mengukur kestabilan dan generalisasi model. Hasil dari setiap fold ditampilkan, serta dihitung rata-rata nilai akurasi dari seluruh fold sebagai gambaran performa umum model.

```
Mean Squared Error (MSE): 0.0339
Cross-validation scores: [0.9485342  0.9257329  0.94723127 0.94332248 0.9380704 ]
Average CV score: 0.9405782502155272
```

Gambar 12. Nilai MSE

KESIMPULAN

Berdasarkan hasil implementasi dan pengujian sistem klasifikasi kualitas air menggunakan metode K-Nearest Neighbor (KNN), dapat disimpulkan bahwa metode ini memberikan performa yang sangat baik dalam mendeteksi kondisi kualitas air. Dengan menggunakan parameter K = 5, model KNN mampu mencapai tingkat akurasi sebesar 97%, yang menunjukkan efektivitasnya dalam mengelompokkan sampel air ke dalam kategori layak atau tidak layak konsumsi. Evaluasi

tambahan melalui confusion matrix, classification report, ROC curve, dan cross-validation turut memperkuat kinerja model yang stabil dan konsisten. Hasil ini membuktikan bahwa KNN dapat digunakan sebagai solusi alternatif dalam sistem monitoring kualitas air secara otomatis, yang berpotensi membantu masyarakat dan instansi terkait dalam mengambil keputusan cepat terhadap potensi kontaminasi air. Dengan demikian, penelitian ini berhasil menunjukkan bahwa metode KNN layak digunakan dalam aplikasi nyata untuk analisis kualitas air berbasis data.

Daftar Pustaka

Aldi, M., Hartono, A., & Saputra, R. (2023). Klasifikasi kualitas air minum menggunakan Naive Bayes, Decision Tree, dan K-Nearest Neighbor. Jurnal Teknologi Informasi, 15(2), 45–52.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Ilyas, M., Nugroho, D., & Ramadhan, F. (2022). Penerapan algoritma Random Forest untuk klasifikasi kualitas air sumur di Jakarta. Jurnal Ilmu Komputer dan Aplikasinya, 10(1), 33–41.

Prismahardi, R., Putra, A., & Yulianto, B. (2023). Analisis kinerja algoritma machine learning dalam klasifikasi kualitas air. Jurnal Sistem Informasi, 12(3), 60–68.

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). Introduction to Data Mining (2nd ed.). Pearson Education.

UC Irvine Machine Learning Repository. (2004). Air Quality Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/Air+Quality